601. CrossMPT: Cross-attention Message-passing Transformer for Error Correcting Codes

链接：https://iclr.cc/virtual/2025/poster/28830 abstract： Error correcting codes (ECCs) are indispensable for reliable transmission in communication systems. Recent advancements in deep learning have catalyzed the exploration of ECC decoders based on neural networks. Among these, transformer-based neural decoders have achieved state-of-the-art decoding performance. In this paper, we propose a novel Cross-Attention Message-Passing Transformer (CrossMPT), which shares key operational principles with conventional message-passing decoders. While conventional transformer-based decoders employ a self-attention mechanism without distinguishing between magnitude and syndrome embeddings, CrossMPT updates these two types of embeddings separately and iteratively via two masked cross-attention blocks. The mask matrices are determined by the code's parity-check matrix, which explicitly captures and removes irrelevant relationships between the magnitude and syndrome embeddings. Our experimental results show that CrossMPT significantly outperforms existing neural network-based decoders for various code classes. Notably, CrossMPT achieves this decoding performance improvement while significantly reducing memory usage, computational complexity, inference time, and training time.

602. 6DGS: Enhanced Direction-Aware Gaussian Splatting for Volumetric Rendering

链接：https://iclr.cc/virtual/2025/poster/28126 abstract： Novel view synthesis has advanced significantly with the development of neural radiance fields (NeRF) and 3D Gaussian splatting (3DGS). However, achieving high quality without compromising real-time rendering remains challenging, particularly for physically-based rendering using ray/path tracing with view-dependent effects. Recently, N-dimensional Gaussians (N-DG) introduced a 6D spatial-angular representation to better incorporate view-dependent effects, but the Gaussian representation and control scheme are sub-optimal. In this paper, we revisit 6D Gaussians and introduce 6D Gaussian Splatting (6DGS), which enhances color and opacity representations and leverages the additional directional information in the 6D space for optimized Gaussian control. Our approach is fully compatible with the 3DGS framework and significantly improves real-time radiance field rendering by better modeling view-dependent effects and fine details. Experiments demonstrate that 6DGS significantly outperforms 3DGS and N-DG, achieving up to a 15.73 dB improvement in PSNR with a reduction of 66.5\% Gaussian points compared to 3DGS. The project page is: https://gaozhongpai.github.io/6dgs/.

603. 3D Vision-Language Gaussian Splatting

链接：https://iclr.cc/virtual/2025/poster/29604 abstract： Recent advancements in 3D reconstruction methods and vision-language models have propelled the development of multi-modal 3D scene understanding, which has vital applications in robotics, autonomous driving, and virtual/augmented reality. However, current multi-modal scene understanding approaches have naively embedded semantic representations into 3D reconstruction methods without striking a balance between visual and language modalities, which leads to unsatisfying semantic rasterization of translucent or reflective objects, as well as over-fitting on color modality. To alleviate these limitations, we propose a solution that adequately handles the distinct visual and semantic modalities, i.e., a 3D vision-language Gaussian splatting model for scene understanding, to put emphasis on the representation learning of language modality. We propose a novel cross-modal rasterizer, using modality fusion along with a smoothed semantic indicator for enhancing semantic rasterization. We also employ a camera-view blending technique to improve semantic consistency between existing and synthesized views, thereby effectively mitigating over-fitting. Extensive experiments demonstrate that our method achieves state-of-the-art performance in open-vocabulary semantic segmentation, surpassing existing methods by a significant margin.

604. Order-aware Interactive Segmentation

链接：https://iclr.cc/virtual/2025/poster/30760 abstract： Interactive segmentation aims to accurately segment target objects with minimal user interactions. However, current methods often fail to accurately separate target objects from the background, due to a limited understanding of order, the relative depth between objects in a scene. To address this issue, we propose OIS: order-aware interactive segmentation, where we explicitly encode the relative depth between objects into order maps. We introduce a novel order-aware attention, where the order maps seamlessly guide the user interactions (in the form of clicks) to attend to the image features. We further present an object-aware attention module to incorporate a strong object-level understanding to better differentiate objects with similar order. Our approach allows both dense and sparse integration of user clicks, enhancing both accuracy and efficiency as compared to prior works. Experimental results demonstrate that OIS achieves state-of-the-art performance, improving mIoU after one click by 7.61 on the HQSeg44K dataset and 1.32 on the DAVIS dataset as compared to the previous state-of-the-art SegNext, while also doubling inference speed compared to current leading methods.

605. New Algorithms for the Learning-Augmented k-means Problem

链接：https://iclr.cc/virtual/2025/poster/29284 abstract： In this paper, we study the clustering problems in the learning-augmented setting, where predicted labels for a d-dimensional dataset with size m are given by an oracle to serve as auxiliary information to improve the clustering performance. Following the prior work, the given oracle is parameterized by some error

rate α, which captures the accuracy of the oracle such that there are at most α fraction of false positives and false negatives in each predicted cluster. In this setting, the goal is to design fast and practical algorithms that can break the computational barriers of inapproximability. The current state-of-the-art learning-augmented k-means algorithm relies on sorting strategies to find good coordinates approximation, where a (1+O(α))-approximation can be achieved with near-linear running time in the data size. However, the computational demands for sorting may limit the scalability of the algorithm for handling large-scale datasets. To address this issue, in this paper, we propose new algorithms that can identify good coordinates approximation using sampling-based strategies, where (1+O(α))-approximation can be achieved with linear running time in the data size. To obtain a more practical algorithm for the problem with better clustering quality and running time, we propose a sampling-based heuristic which can directly find center approximations using sampling-based strategies. Empirical experiments show that our proposed methods are faster than the state-of-the-art learning-augmented k-means algorithms with comparable performances on clustering quality.

606. Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG

链接：https://iclr.cc/virtual/2025/poster/28355 abstract： Retrieval-augmented generation (RAG) empowers large language models (LLMs) to utilize external knowledge sources. The increasing capacity of LLMs to process longer input sequences opens up avenues for providing more retrieved information, to potentially enhance the quality of generated outputs. From a long-context LLM perspective, it assumes that a larger retrieval set would contain more relevant information (higher recall), that might result in improved performance. However, our empirical findings demonstrate that for many long-context LLMs, the quality of generated output initially improves first, but then subsequently declines as the number of retrieved passages increases. This paper investigates this phenomenon, identifying the detrimental impact of retrieved "hard negatives" as a key contributor. To mitigate this and enhance the robustness of long-context LLM-based RAG, we propose both training-free and training-based approaches. We first showcase the effectiveness of retrieval reordering as a simple yet powerful training-free optimization. Furthermore, we explore training-based methods, specifically RAG-specific implicit LLM fine-tuning and RAG-oriented fine-tuning with intermediate reasoning, demonstrating their capacity for substantial performance gains. Finally, we conduct a systematic analysis of design choices for these training-based methods, including data distribution, retriever selection, and training context length.

607. RAG-DDR: Optimizing Retrieval-Augmented Generation Using Differentiable Data Rewards

链接：https://iclr.cc/virtual/2025/poster/29729 abstract： Retrieval-Augmented Generation (RAG) has proven its effectiveness in mitigating hallucinations in Large Language Models (LLMs) by retrieving knowledge from external resources. To adapt LLMs for the RAG systems, current approaches use instruction tuning to optimize LLMs, improving their ability to utilize retrieved knowledge. This supervised fine-tuning (SFT) approach focuses on equipping LLMs to handle diverse RAG tasks using different instructions. However, it trains RAG modules to overfit training signals and overlooks the varying data preferences among agents within the RAG system. In this paper, we propose a Differentiable Data Rewards (DDR) method, which end-to-end trains RAG systems by aligning data preferences between different RAG modules. DDR works by collecting the rewards to optimize each agent in the RAG system with the rollout method, which prompts agents to sample some potential responses as perturbations, evaluates the impact of these perturbations on the whole RAG system, and subsequently optimizes the agent to produce outputs that improve the performance of the RAG system. Our experiments on various knowledge-intensive tasks demonstrate that DDR significantly outperforms the SFT method, particularly for LLMs with smaller-scale parameters that depend more on the retrieved knowledge. Additionally, DDR exhibits a stronger capability to align the data preference between RAG modules. The DDR method makes the generation module more effective in extracting key information from documents and mitigating conflicts between parametric memory and external knowledge. All codes are available at https://github.com/OpenMatch/RAG-DDR.

608. Discovering Temporally Compositional Neural Manifolds with Switching Infinite GPFA

链接：https://iclr.cc/virtual/2025/poster/31122 abstract： Gaussian Process Factor Analysis (GPFA) is a powerful latent variable model for extracting low-dimensional manifolds underlying population neural activities. However, one limitation of standard GPFA models is that the number of latent factors needs to be pre-specified or selected through heuristic-based processes, and that all factors contribute at all times. We propose the infinite GPFA model, a fully Bayesian non-parametric extension of the classical GPFA by incorporating an Indian Buffet Process (IBP) prior over the factor loading process, such that it is possible to infer a potentially infinite set of latent factors, and the identity of those factors that contribute to neural firings in a compositional manner at \textit{each} time point. Learning and inference in the infinite GPFA model is performed through variational expectation-maximisation, and we additionally propose scalable extensions based on sparse variational Gaussian Process methods. We empirically demonstrate that the infinite GPFA model correctly infers dynamically changing activations of latent factors on a synthetic dataset. By fitting the infinite GPFA model to population activities of hippocampal place cells during spatial tasks with alternating random foraging and spatial memory phases, we identify novel non-trivial and behaviourally meaningful dynamics in the neural encoding process.

609. ZIP: An Efficient Zeroth-order Prompt Tuning for Black-box Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/31142 abstract： Recent studies have introduced various approaches for prompt-tuning black-box vision-language models, referred to as black-box prompt-tuning (BBPT). While BBPT has demonstrated considerable potential, it is often found that many existing methods require an excessive number of queries (i.e., function evaluations), which poses a significant challenge in real-world scenarios where the number of allowed queries is limited. To tackle this issue, we propose Zeroth-order Intrinsic-dimensional Prompt-tuning (ZIP), a novel approach that enables efficient and robust prompt optimization in a purely black-box setting. The key idea of ZIP is to reduce the problem dimensionality and the variance of zeroth-order gradient estimates, such that the training is done fast with far less queries. We achieve this by re-parameterizing prompts in low-rank representations and designing intrinsic-dimensional clipping of estimated gradients. We evaluate ZIP on 13+ vision-language tasks in standard benchmarks and show that it achieves an average improvement of approximately 6% in few-shot accuracy and 48% in query efficiency compared to the best-performing alternative BBPT methods, establishing a new state of the art. Our ablation analysis further shows that the proposed clipping mechanism is robust and nearly optimal, without the need to manually select the clipping threshold, matching the result of expensive hyperparameter search.

610. Unlearning-based Neural Interpretations

链接：https://iclr.cc/virtual/2025/poster/29768 abstract： Gradient-based interpretations often require an anchor point of comparison to avoid saturation in computing feature importance. We show that current baselines defined using static functions—constant mapping, averaging or blurring—inject harmful colour, texture or frequency assumptions that deviate from model behaviour. This leads to accumulation of irregular gradients, resulting in attribution maps that are biased, fragile and manipulable. Departing from the static approach, we propose $\texttt{UNI}$ to compute an (un)learnable, debiased and adaptive baseline by perturbing the input towards an $\textit{unlearning direction}$ of steepest ascent. Our method discovers reliable baselines and succeeds in erasing salient features, which in turn locally smooths the high-curvature decision boundaries. Our analyses point to unlearning as a promising avenue for generating faithful, efficient and robust interpretations.

611. Bridging Information Asymmetry in Text-video Retrieval: A Data-centric Approach

链接：https://iclr.cc/virtual/2025/poster/29517 abstract： As online video content rapidly grows, the task of text-video retrieval (TVR) becomes increasingly important. A key challenge in TVR is the information asymmetry between video and text: videos are inherently richer in information, while their textual descriptions often capture only fragments of this complexity. This paper introduces a novel, data-centric framework to bridge this gap by enriching textual representations to better match the richness of video content. During training, videos are segmented into event-level clips and captioned to ensure comprehensive coverage. During retrieval, a large language model (LLM) generates semantically diverse queries to capture a broader range of possible matches. To enhance retrieval efficiency, we propose a query selection mechanism that identifies the most relevant and diverse queries, reducing computational cost while improving accuracy. Our method achieves state-of-the-art results across multiple benchmarks, demonstrating the power of data-centric approaches in addressing information asymmetry in TVR. This work paves the way for new research focused on leveraging data to improve cross-modal retrieval.

612. NovelQA: Benchmarking Question Answering on Documents Exceeding 200K Tokens

链接：https://iclr.cc/virtual/2025/poster/27983 abstract： Recent advancements in Large Language Models (LLMs) have pushed the boundaries of natural language processing, especially in long-context understanding. However, the evaluation of these models' long-context abilities remains a challenge due to the limitations of current benchmarks. To address this gap, we introduce NovelQA, a benchmark tailored for evaluating LLMs with complex, extended narratives. NovelQA, constructed from English novels, offers a unique blend of complexity, length, and narrative coherence, making it an ideal tool for assessing deep textual understanding in LLMs. This paper details the design and construction of NovelQA, focusing on its comprehensive manual annotation process and the variety of question types aimed at evaluating nuanced comprehension. Our evaluation of long-context LLMs on NovelQA reveals significant insights into their strengths and weaknesses. Notably, the models struggle with multi-hop reasoning, detail-oriented questions, and handling extremely long inputs, averaging over 200,000 tokens. Results highlight the need for substantial advancements in LLMs to enhance their long-context comprehension and contribute effectively to computational literary analysis.

613. A Truncated Newton Method for Optimal Transport

链接：https://iclr.cc/virtual/2025/poster/28814 abstract： Developing a contemporary optimal transport (OT) solver requires navigating trade-offs among several critical requirements: GPU parallelization, scalability to high-dimensional problems, theoretical convergence guarantees, empirical performance in terms of precision versus runtime, and numerical stability in practice. With these challenges in mind, we introduce a specialized truncated Newton algorithm for entropic-regularized OT. In addition to proving that locally quadratic convergence is possible without assuming a Lipschitz Hessian, we provide strategies to maximally exploit the high rate of local convergence in practice. Our GPU-parallel algorithm exhibits exceptionally favorable

runtime performance, achieving high precision orders of magnitude faster than many existing alternatives. This is evidenced by wall-clock time experiments on 24 problem sets (12 datasets $\times$ 2 cost functions). The scalability of the algorithm is showcased on an extremely large OT problem with $n \approx 10^6$, solved approximately under weak entopric regularization.

614. MamBEV: Enabling State Space Models to Learn Birds-Eye-View Representations

链接：https://iclr.cc/virtual/2025/poster/29911 abstract： 3D visual perception tasks, such as 3D detection from multi-camera images, are essential components of autonomous driving and assistance systems. However, designing computationally efficient methods remains a significant challenge. In this paper, we propose a Mamba-based framework called MamBEV, which learns unified Bird's Eye View (BEV) representations using linear spatio-temporal SSM-based attention. This approach supports multiple 3D perception tasks with significantly improved computational and memory efficiency. Furthermore, we introduce SSM based cross-attention, analogous to standard cross attention, where BEV query representations can interact with relevant image features. Extensive experiments demonstrate MamBEV's promising performance across diverse visual perception metrics, highlighting its advantages in input scaling efficiency compared to existing benchmark models.

615. Sharpness-Aware Minimization Efficiently Selects Flatter Minima Late In Training

链接：https://iclr.cc/virtual/2025/poster/29182 abstract： Sharpness-Aware Minimization (SAM) has substantially improved the generalization of neural networks under various settings. Despite the success, its effectiveness remains poorly understood. In this work, we discover an intriguing phenomenon in the training dynamics of SAM, shedding light on understanding its implicit bias towards flatter minima over Stochastic Gradient Descent (SGD). Specifically, we find that SAM efficiently selects flatter minima late in training. Remarkably, even a few epochs of SAM applied at the end of training yield nearly the same generalization and solution sharpness as full SAM training. Subsequently, we delve deeper into the underlying mechanism behind this phenomenon. Theoretically, we identify two phases in the learning dynamics after applying SAM late in training: i) SAM first escapes the minimum found by SGD exponentially fast; and ii) then rapidly converges to a flatter minimum within the same valley. Furthermore, we empirically investigate the role of SAM during the early training phase. We conjecture that the optimization method chosen in the late phase is more crucial in shaping the final solution's properties. Based on this viewpoint, we extend our findings from SAM to Adversarial Training.

616. Do Deep Neural Network Solutions Form a Star Domain?

链接：https://iclr.cc/virtual/2025/poster/29678 abstract： It has recently been conjectured that neural network solution sets reachable via stochastic gradient descent (SGD) are convex, considering permutation invariances. This means that a linear path can connect two independent solutions with low loss, given the weights of one of the models are appropriately permuted. However, current methods to test this theory often require very wide networks to succeed. In this work, we conjecture that more generally, the SGD solution set is a star domain that contains a star model that is linearly connected to all the other solutions via paths with low loss values, modulo permutations. We propose the Starlight algorithm that finds a star model of a given learning task. We validate our claim by showing that this star model is linearly connected with other independently found solutions. As an additional benefit of our study, we demonstrate better uncertainty estimates on Bayesian Model Averaging over the obtained star domain. Further, we demonstrate star models as potential substitutes for model ensembles.

617. On the Optimization and Generalization of Two-layer Transformers with Sign Gradient Descent

链接：https://iclr.cc/virtual/2025/poster/30719 abstract： The Adam optimizer is widely used for transformer optimization in practice, which makes understanding the underlying optimization mechanisms an important problem.However, due to the Adam's complexity, theoretical analysis of how it optimizes transformers remains a challenging task. Fortunately, Sign Gradient Descent (SignGD) serves as an effective surrogate for Adam.Despite its simplicity, theoretical understanding of how SignGD optimizes transformers still lags behind.In this work, we study how SignGD optimizes a two-layer transformer -- consisting of a softmax attention layer with trainable query-key parameterization followed by a linear layer -- on a linearly separable noisy dataset.We identify four stages in the training dynamics, each exhibiting intriguing behaviors.Based on the training dynamics, we prove the fast convergence but poor generalization of the learned transformer on the noisy dataset.We also show that Adam behaves similarly to SignGD in terms of both optimization and generalization in this setting.Additionally, we find that the poor generalization of SignGD is not solely due to data noise,suggesting that both SignGD and Adam requires high-quality data for real-world tasks.Finally, experiments on synthetic and real-world datasets empirically support our theoretical results.

618. Personalized Representation from Personalized Generation

链接：https://iclr.cc/virtual/2025/poster/32064 abstract： Modern vision models excel at general purpose downstream tasks. It is unclear, however, how they may be used for personalized vision tasks, which are both fine-grained and data-scarce. Recent works have successfully applied synthetic data to general-purpose representation learning, while advances in T2I diffusion models have enabled the generation of personalized images from just a few real examples. Here, we explore a potential

connection between these ideas, and formalize the challenge of using personalized synthetic data to learn personalized representations, which encode knowledge about an object of interest and may be flexibly applied to any downstream task relating to the target object. We introduce an evaluation suite for this challenge, including reformulations of two existing datasets and a novel dataset explicitly constructed for this purpose, and propose a contrastive learning approach that makes creative use of image generators. We show that our method improves personalized representation learning for diverse downstream tasks, from recognition to segmentation, and analyze characteristics of image generation approaches that are key to this gain.

619. Beyond Circuit Connections: A Non-Message Passing Graph Transformer Approach for Quantum Error Mitigation

链接：https://iclr.cc/virtual/2025/poster/29291 abstract： Despite the progress in quantum computing, one major bottleneck against the practical utility is its susceptibility to noise, which frequently occurs in current quantum systems. Existing quantum error mitigation (QEM) methods either lack generality to noise and circuit types or fail to capture the global dependencies of entire systems in addition to circuit structure. In this work, we first propose a unique circuit-to-graph encoding scheme with qubit-wise noisy measurement aggregated. Then, we introduce GTranQEM, a non-message passing graph transformer designed to mitigate errors in expected circuit measurement outcomes effectively. GTranQEM is equipped with a quantum-specific positional encoding, a structure matrix as attention bias guiding nonlocal aggregation, and a virtual quantum-representative node to further grasp graph representations, which guarantees to model the long-range entanglement. Experimental evaluations demonstrate that GTranQEM outperforms state-of-the-art QEM methods on both random and structured quantum circuits across noise types and scales among diverse settings.

620. Self-Introspective Decoding: Alleviating Hallucinations for Large Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/28166 abstract： Hallucination remains a significant challenge in Large Vision-Language Models (LVLMs). To alleviate this issue, some methods, known as contrastive decoding, induce hallucinations by manually disturbing the raw vision or instruction inputs and then mitigate them by contrasting the outputs of the original and disturbed LVLMs. However, these holistic input disturbances sometimes induce potential noise and also double the inference cost. To tackle these issues, we propose a simple yet effective method named $\textit{Self-Introspective Decoding}$ (SID). Our empirical investigations reveal that pre-trained LVLMs can introspectively assess the importance of vision tokens based on preceding vision and text (both instruction and generated) tokens. Leveraging this insight, we develop the Context and Text-aware Token Selection (CT^{2}S) strategy, which preserves only the least important vision tokens after the early decoder layers, thereby adaptively amplify vision-and-text association hallucinations during auto-regressive decoding. This strategy ensures that multimodal knowledge absorbed in the early decoder layers induces multimodal contextual rather than aimless hallucinations, and significantly reduces computation burdens. Subsequently, the original token logits subtract the amplified fine-grained hallucinations, effectively alleviating hallucinations without compromising the LVLMs' general ability. Extensive experiments illustrate SID generates less-hallucination and higher-quality texts across various metrics, without much additional computation cost.

621. ToolACE: Winning the Points of LLM Function Calling

链接：https://iclr.cc/virtual/2025/poster/30779 abstract： Function calling significantly extends the application boundary of large language models (LLMs), where high-quality and diverse training data is critical for unlocking this capability. However, collecting and annotating real function-calling data is challenging, while synthetic data from existing pipelines often lack coverage and accuracy. In this paper, we present ToolACE, an automatic agentic pipeline designed to generate accurate, complex, and diverse tool-learning data, specifically tailored to the capabilities of LLMs. ToolACE leverages a novel self-evolution synthesis process to curate a comprehensive API pool of 26,507 diverse APIs. Dialogs are further generated through the interplay among multiple agents, under the guidance of a complexity evaluator. To ensure data accuracy, we implement a dual-layer verification system combining rule-based and model-based checks. We demonstrate that models trained on our synthesized data---even with only 8B parameters---achieve state-of-the-art performance, comparable to the latest GPT-4 models. Our model and a subset of the data are publicly available at https://huggingface.co/Team-ACE.

622. Differentially private optimization for non-decomposable objective functions

链接：https://iclr.cc/virtual/2025/poster/30368 abstract： Unsupervised pre-training is a common step in developing computer vision models and large language models. In this setting, the absence of labels requires the use of similarity-based loss functions, such as the contrastive loss, that favor minimizing the distance between similar inputs and maximizing the distance between distinct inputs. As privacy concerns mount, training these models using differential privacy has become more important. However, due to how inputs are generated for these losses, one of their undesirable properties is that their L_2 sensitivity grows with the batch size. This property is particularly disadvantageous for differentially private training methods, such as DP-SGD. To overcome this issue, we develop a new DP-SGD variant for similarity based loss functions --- in particular, the commonly-used contrastive loss --- that manipulates gradients of the objective function in a novel way to obtain a sensitivity of the summed gradient that is $O(1)$ for batch size n. We test our DP-SGD variant on some CIFAR-10 pre-training and CIFAR-

100 finetuning tasks and show that, in both tasks, our method's performance comes close to that of a non-private model and generally outperforms DP-SGD applied directly to the contrastive loss.

623. Multi-Robot Motion Planning with Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/30630 abstract： Diffusion models have recently been successfully applied to a wide range of robotics applications for learning complex multi-modal behaviors from data. However, prior works have mostly been confined to single-robot and small-scale environments due to the high sample complexity of learning multi-robot diffusion models. In this paper, we propose a method for generating collision-free multi-robot trajectories that conform to underlying data distributions while using only single-robot data. Our algorithm, Multi-robot Multi-model planning Diffusion (MMD), does so by combining learned diffusion models with classical search-based techniques---generating data-driven motions under collision constraints. Scaling further, we show how to compose multiple diffusion models to plan in large environments where a single diffusion model fails to generalize well. We demonstrate the effectiveness of our approach in planning for dozens of robots in a variety of simulated scenarios motivated by logistics environments.

624. Beyond Graphs: Can Large Language Models Comprehend Hypergraphs?

链接：https://iclr.cc/virtual/2025/poster/31158 abstract： Existing benchmarks like NLGraph and GraphQA evaluate LLMs on graphs by focusing mainly on pairwise relationships, overlooking the high-order correlations found in real-world data. Hypergraphs, which can model complex beyond-pairwise relationships, offer a more robust framework but are still underexplored in the context of LLMs. To address this gap, we introduce LLM4Hypergraph, the first comprehensive benchmark comprising 21,500 problems across eight low-order, five high-order, and two isomorphism tasks, utilizing both synthetic and real-world hypergraphs from citation networks and protein structures. We evaluate six prominent LLMs, including GPT-4o, demonstrating our benchmark’s effectiveness in identifying model strengths and weaknesses. Our specialized prompt- ing framework incorporates seven hypergraph languages and introduces two novel techniques, Hyper-BAG and Hyper-COT, which enhance high-order reasoning and achieve an average 4% (up to 9%) performance improvement on structure classification tasks. This work establishes a foundational testbed for integrating hypergraph computational capabilities into LLMs, advancing their comprehension.

625. ImDy: Human Inverse Dynamics from Imitated Observations

链接：https://iclr.cc/virtual/2025/poster/29085 abstract： Inverse dynamics (ID), which aims at reproducing the driven torques from human kinematic observations, has been a critical tool for gait analysis. However, it is hindered from wider application to general motion due to its limited scalability. Conventional optimization-based ID requires expensive laboratory setups, restricting its availability. To alleviate this problem, we propose to exploit the recently progressive human motion imitation algorithms to learn human inverse dynamics in a data-driven manner. The key insight is that the human ID knowledge is implicitly possessed by motion imitators, though not directly applicable. In light of this, we devise an efficient data collection pipeline with state-of-the-art motion imitation algorithms and physics simulators, resulting in a large-scale human inverse dynamics benchmark as Imitated Dynamics (ImDy). ImDy contains over 150 hours of motion with joint torque and full-body ground reaction force data. With ImDy, we train a data-driven human inverse dynamics solver ImDyS(olver) in a fully supervised manner, which conducts ID and ground reaction force estimation simultaneously. Experiments on ImDy and real-world data demonstrate the impressive competency of ImDyS in human inverse dynamics and ground reaction force estimation. Moreover, the potential of ImDy(-S) as a fundamental motion analysis tool is exhibited with downstream applications. The project page is https://foruck.github.io/ImDy.

626. HarmAug: Effective Data Augmentation for Knowledge Distillation of Safety Guard Models

链接：https://iclr.cc/virtual/2025/poster/27754 abstract： Safety guard models that detect malicious queries aimed at large language models (LLMs) are essential for ensuring the secure and responsible deployment of LLMs in real-world applications.However, deploying existing safety guard models with billions of parameters alongside LLMs on mobile devices is impractical due to substantial memory requirements and latency.To reduce this cost, we distill a large teacher safety guard model into a smaller one using a labeled dataset of instruction-response pairs with binary harmfulness labels. Due to the limited diversity of harmful instructions in the existing labeled dataset, naively distilled models tend to underperform compared to larger models. To bridge the gap between small and large models, we propose HarmAug, a simple yet effective data augmentation method that involves jailbreaking an LLM and prompting it to generate harmful instructions. Given a prompt such as, "Make a single harmful instruction prompt that would elicit offensive content", we add an affirmative prefix (e.g., "I have an idea for a prompt:") to the LLM's response. This encourages the LLM to continue generating the rest of the response, leading to sampling harmful instructions. Another LLM generates a response to the harmful instruction, and the teacher model labels the instruction-response pair. We empirically show that our HarmAug outperforms other relevant baselines. Moreover, a 435-million-parameter safety guard model trained with HarmAug achieves an F1 score comparable to larger models with over 7 billion parameters, and even outperforms them in AUPRC, while operating at less than 25\% of their computational cost. Our code, safety guard model, and synthetic dataset are publicly available.

627. ReMatching Dynamic Reconstruction Flow

链接：https://iclr.cc/virtual/2025/poster/29079 abstract： Reconstructing a dynamic scene from image inputs is a fundamental computervision task with many downstream applications. Despite recent advancements, existing approaches still struggle to achieve high-quality reconstructions from unseen viewpoints and timestamps. This work introduces the ReMatching framework, designed to improve reconstruction quality by incorporating deformation priors into dynamic reconstruction models. Our approach advocates for velocity-field based priors, for which we suggest a matching procedure that can seamlessly supplement existing dynamic reconstruction pipelines. The framework is highly adaptable and can be applied to various dynamic representations. Moreover, it supports integrating multiple types of model priors and enables combining simpler ones to create more complex classes. Our evaluations on popular benchmarks involving both synthetic and real-world dynamic scenes demonstrate that augmenting current state-of-the-art methods with our approach leads to a clear improvement in reconstruction accuracy.

628. Feature Averaging: An Implicit Bias of Gradient Descent Leading to Non-Robustness in Neural Networks

链接：https://iclr.cc/virtual/2025/poster/27668 abstract： In this work, we investigate a particular implicit bias in gradient descent training, which we term “Feature Averaging,” and argue that it is one of the principal factors contributing to the non-robustness of deep neural networks. We show that, even when multiple discriminative features are present in the input data, neural networks trained by gradient descent tend to rely on an average (or a certain combination) of these features for classification, rather than distinguishing and leveraging each feature individually. Specifically, we provide a detailed theoretical analysis of the training dynamics of two-layer ReLU networks on a binary classification task, where the data distribution consists of multiple clusters with mutually orthogonal centers. We rigorously prove that gradient descent biases the network towards feature averaging, where the weights of each hidden neuron represent an average of the cluster centers (each corresponding to a distinct feature), thereby making the network vulnerable to input perturbations aligned with the negative direction of the averaged features. On the positive side, we demonstrate that this vulnerability can be mitigated through more granular supervision. In particular, we prove that a two-layer ReLU network can achieve optimal robustness when trained to classify individual features rather than merely the original binary classes. Finally, we validate our theoretical findings with experiments on synthetic datasets, MNIST, and CIFAR-10, and confirm the prevalence of feature averaging and its impact on adversarial robustness. We hope these theoretical and empirical insights deepen the understanding of how gradient descent shapes feature learning and adversarial robustness, and how more detailed supervision can enhance robustness.

629. Efficient Exploration and Discriminative World Model Learning with an Object-Centric Abstraction

链接：https://iclr.cc/virtual/2025/poster/28750 abstract： In the face of difficult exploration problems in reinforcement learning, we study whether giving an agent an object-centric mapping (describing a set of items and their attributes) allow for more efficient learning. We found this problem is best solved hierarchically by modelling items at a higher level of state abstraction to pixels, and attribute change at a higher level of temporal abstraction to primitive actions. This abstraction simplifies the transition dynamic by making specific future states easier to predict. We make use of this to propose a fully model-based algorithm that learns a discriminative world model, plans to explore efficiently with only a count-based intrinsic reward, and can subsequently plan to reach any discovered (abstract) states.We demonstrate the model's ability to (i) efficiently solve single tasks, (ii) transfer zero-shot and few-shot across item types and environments, and (iii) plan across long horizons. Across a suite of 2D crafting and MiniHack environments, we empirically show our model significantly out-performs state-of-the-art low-level methods (without abstraction), as well as performant model-free and model-based methods using the same abstraction. Finally, we show how to learn low level object-perturbing policies via reinforcement learning, and the object mapping itself by supervised learning.

630. HOPE for a Robust Parameterization of Long-memory State Space Models

链接：https://iclr.cc/virtual/2025/poster/29646 abstract： State-space models (SSMs) that utilize linear, time-invariant (LTI) systems are known for their effectiveness in learning long sequences. To achieve state-of-the-art performance, an SSM often needs a specifically designed initialization, and the training of state matrices is on a logarithmic scale with a very small learning rate. To understand these choices from a unified perspective, we view SSMs through the lens of Hankel operator theory. Building upon it, we develop a new parameterization scheme, called HOPE, for LTI systems that utilizes Markov parameters within Hankel operators. Our approach helps improve the initialization and training stability, leading to a more robust parameterization. We efficiently implement these innovations by nonuniformly sampling the transfer functions of LTI systems, and they require fewer parameters compared to canonical SSMs. When benchmarked against HiPPO-initialized models such as S4 and S4D, an SSM parameterized by Hankel operators demonstrates improved performance on Long-Range Arena (LRA) tasks. Moreover, our new parameterization endows the SSM with non-decaying memory within a fixed time window, which is empirically corroborated by a sequential CIFAR-10 task with padded noise.

631. CFD: Learning Generalized Molecular Representation via Concept-

Enhanced Feedback Disentanglement

链接：https://iclr.cc/virtual/2025/poster/30491 abstract： To accelerate biochemical research, e.g., drug and protein discovery, molecular representation learning (MRL) has attracted much attention. However, most existing methods follow the closed-set assumption that training and testing data share identical distribution, which limits their generalization abilities in out-of-distribution (OOD) cases. In this paper, we explore designing a new disentangled mechanism for learning generalized molecular representation that exhibits robustness against distribution shifts. And an approach of Concept-Enhanced Feedback Disentanglement (CFD) is proposed, whose goal is to exploit the feedback mechanism to learn distribution-agnostic representation. Specifically, we first propose two dedicated variational encoders to separately decompose distribution-agnostic and spurious features. Then, a set of molecule-aware concepts are tapped to focus on invariant substructure characteristics. By fusing these concepts into the disentangled distribution-agnostic features, the generalization ability of the learned molecular representation could be further enhanced. Next, we execute iteratively the disentangled operations based on a feedback received from the previous output. Finally, based on the outputs of multiple feedback iterations, we construct a self-supervised objective to promote the variational encoders to possess the disentangled capability. In the experiments, our method is verified on multiple real-world molecular datasets. The significant performance gains over state-of-the-art baselines demonstrate that our method can effectively disentangle generalized molecular representation in the presence of various distribution shifts. The source code will be released at https://github.com/AmingWu/MoleculeCFD.

632. Analysis of Linear Mode Connectivity via Permutation-Based Weight Matching: With Insights into Other Permutation Search Methods

链接：https://iclr.cc/virtual/2025/poster/28512 abstract： Recently, Ainsworth et al. (2023) showed that using weight matching (WM) to minimize the L^2 distance in a permutation search of model parameters effectively identifies permutations that satisfy linear mode connectivity (LMC), where the loss along a linear path between two independently trained models with different seeds remains nearly constant. This paper analyzes LMC using WM, which is useful for understanding stochastic gradient descent's effectiveness and its application in areas like model merging. We first empirically show that permutations found by WM do not significantly reduce the L^2 distance between two models, and the occurrence of LMC is not merely due to distance reduction by WM itself. We then demonstrate that permutations can change the directions of the singular vectors, but not the singular values, of the weight matrices in each layer. This finding shows that permutations found by WM primarily align the directions of singular vectors associated with large singular values across models. This alignment brings the singular vectors with large singular values, which determine the model's functionality, closer between the original and merged models, allowing the merged model to retain functionality similar to the original models, thereby satisfying LMC. This paper also analyzes activation matching (AM) in terms of singular vectors and finds that the principle of AM is likely the same as that of WM. Finally, we analyze the difference between WM and the straight-through estimator (STE), a dataset-dependent permutation search method, and show that WM can be more advantageous than STE in achieving LMC among three or more models.

633. Scalable Benchmarking and Robust Learning for Noise-Free Ego-Motion and 3D Reconstruction from Noisy Video

链接：https://iclr.cc/virtual/2025/poster/29719 abstract： We aim to redefine robust ego-motion estimation and photorealistic 3D reconstruction by addressing a critical limitation: the reliance on noise-free data in existing models. While such sanitized conditions simplify evaluation, they fail to capture the unpredictable, noisy complexities of real-world environments. Dynamic motion, sensor imperfections, and synchronization perturbations lead to sharp performance declines when these models are deployed in practice, revealing an urgent need for frameworks that embrace and excel under real-world noise.To bridge this gap, we tackle three core challenges: scalable data generation, comprehensive benchmarking, and model robustness enhancement. First, we introduce a scalable noisy data synthesis pipeline that generates diverse datasets simulating complex motion, sensor imperfections, and synchronization errors. Second, we leverage this pipeline to create Robust-Ego3D, a benchmark rigorously designed to expose noise-induced performance degradation, highlighting the limitations of current learning-based methods in ego-motion accuracy and 3D reconstruction quality. Third, we propose Correspondence-guided Gaussian Splatting (CorrGS), a novel method that progressively refines an internal clean 3D representation by aligning noisy observations with rendered RGB-D frames from clean 3D map, enhancing geometric alignment and appearance restoration through visual correspondence.Extensive experiments on synthetic and real-world data demonstrate that CorrGS consistently outperforms prior state-of-the-art methods, particularly in scenarios involving rapid motion and dynamic illumination. We will release our code and benchmark to advance robust 3D vision, setting a new standard for ego-motion estimation and high-fidelity reconstruction in noisy environments.

634. DoF: A Diffusion Factorization Framework for Offline Multi-Agent Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29825 abstract： Diffusion models have been widely adopted in image and language generation and are now being applied to reinforcement learning. However, the application of diffusion models in offline cooperative Multi-Agent Reinforcement Learning (MARL) remains limited. Although existing studies explore this direction, they suffer from scalability or poor cooperation issues due to the lack of design principles for diffusion-based MARL. The Individual-Global-Max (IGM) principle is a popular design principle for cooperative MARL. By satisfying this principle, MARL algorithms

achieve remarkable performance with good scalability. In this work, we extend the IGM principle to the Individual-Global-identically-Distributed (IGD) principle. This principle stipulates that the generated outcome of a multi-agent diffusion model should be identically distributed as the collective outcomes from multiple individual-agent diffusion models. We propose DoF, a diffusion factorization framework for Offline MARL. It uses noise factorization function to factorize a centralized diffusion model into multiple diffusion models. We theoretically show that the noise factorization functions satisfy the IGD principle. Furthermore, DoF uses data factorization function to model the complex relationship among data generated by multiple diffusion models. Through extensive experiments, we demonstrate the effectiveness of DoF. The source code is available at https://github.com/xmu-rl-3dv/DoF.

635. THE ROBUSTNESS OF DIFFERENTIABLE CAUSAL DISCOVERY IN MISSPECIFIED SCENARIOS

链接：https://iclr.cc/virtual/2025/poster/28686 abstract： Causal discovery aims to learn causal relationships between variables from targeted data, making it a fundamental task in machine learning. However, causal discovery algorithms often rely on unverifiable causal assumptions, which are usually difficult to satisfy in real-world data, thereby limiting the broad application of causal discovery in practical scenarios. Inspired by these considerations, this work extensively benchmarks the empirical performance of various mainstream causal discovery algorithms, which assume i.i.d. data, under eight model assumption violations. Our experimental results show that differentiable causal discovery methods exhibit robustness under the metrics of Structural Hamming Distance and Structural Intervention Distance of the inferred graphs in commonly used challenging scenarios, except for scale variation. We also provide the theoretical explanations for the performance of differentiable causal discovery methods. Finally, our work aims to comprehensively benchmark the performance of recent differentiable causal discovery methods under model assumption violations, and provide the standard for reasonable evaluation of causal discovery, as well as to further promote its application in real-world scenarios.

636. VL-ICL Bench: The Devil in the Details of Multimodal In-Context Learning

链接：https://iclr.cc/virtual/2025/poster/29026 abstract： Large language models (LLMs) famously exhibit emergent in-context learning (ICL) - the ability to rapidly adapt to new tasks using few-shot examples provided as a prompt, without updating the model's weights. Built on top of LLMs, vision large language models (VLLMs) have advanced significantly in areas such as recognition, reasoning, and grounding. However, investigations into multimodal ICL have predominantly focused on few-shot visual question answering (VQA), and image captioning, which we will show neither exploit the strengths of ICL, nor test its limitations. The broader capabilities and limitations of multimodal ICL remain under-explored. In this study, we introduce a comprehensive benchmark VL-ICL Bench for multimodal in-context learning, encompassing a broad spectrum of tasks that involve both images and text as inputs and outputs, and different types of challenges, from {perception to reasoning and long context length}. We evaluate the abilities of state-of-the-art VLLMs against this benchmark suite, revealing their diverse strengths and weaknesses, and showing that even the most advanced models, such as GPT-4, find the tasks challenging. By highlighting a range of new ICL tasks, and the associated strengths and limitations of existing models, we hope that our dataset will inspire future work on enhancing the in-context learning capabilities of VLLMs, as well as inspire new applications that leverage VLLM ICL. Project page is at https://ys-zong.github.io/VL-ICL/

637. MIND: Math Informed syNthetic Dialogues for Pretraining LLMs

链接：https://iclr.cc/virtual/2025/poster/29504 abstract： The utility of synthetic data to enhance pretraining data quality and hence to improve downstream task accuracy has been widely explored in recent large language models (LLMs). Yet, these approaches fall inadequate in complex, multi-hop and mathematical reasoning tasks as the synthetic data typically fails to add complementary knowledge to the existing raw corpus. In this work, we propose a novel large-scale and diverse Math Informed syNthetic Dialogue (MIND) generation method that improves the mathematical reasoning ability of LLMs. Specifically, using MIND, we generate synthetic conversations based on OpenWebMath (OWM), resulting in a new math corpus, MIND-OWM. Our experiments with different conversational settings reveal that incorporating knowledge gaps between dialog participants is essential for generating high-quality math data. We further identify an effective way to format and integrate synthetic and raw data during pretraining to maximize the gain in mathematical reasoning, emphasizing the need to restructure raw data rather than use it as-is. Compared to pretraining just on raw data, a model pretrained on MIND-OWM shows significant boost in mathematical reasoning (GSM8K: +13.42%, MATH: +2.30%), including superior performance in specialized knowledge (MMLU: +4.55%, MMLU-STEM: +4.28%) and generalpurpose reasoning tasks (GENERAL REASONING: +2.51%).

638. Building Math Agents with Multi-Turn Iterative Preference Learning

链接：https://iclr.cc/virtual/2025/poster/29347 abstract： Recent studies have shown that large language models' (LLMs) mathematical problem-solving capabilities can be enhanced by integrating external tools, such as code interpreters, and employing multi-turn Chain-of-Thought (CoT) reasoning. While current methods focus on synthetic data generation and Supervised Fine-Tuning (SFT), this paper studies the complementary direct preference learning approach to further improve model performance. However, existing direct preference learning algorithms are originally designed for the single-turn chat task, and do not fully address the complexities of multi-turn reasoning and external tool integration required for tool-integrated mathematical reasoning tasks. To fill in this gap, we introduce a multi-turn direct preference learning framework, tailored for this

context, that leverages feedback from code interpreters and optimizes trajectory-level preferences. This framework includes multi-turn DPO and multi-turn KTO as specific implementations. The effectiveness of our framework is validated through training of various language models using an augmented prompt set from the GSM8K and MATH datasets. Our results demonstrate substantial improvements: a supervised fine-tuned Gemma-1.1-it-7B model's performance increased from 77.5% to 83.9% on GSM8K and from 46.1% to 51.2% on MATH. Similarly, a Gemma-2-it-9B model improved from 84.1% to 86.3% on GSM8K and from 51.0% to 54.5% on MATH.

639. Towards Robust Multimodal Open-set Test-time Adaptation via Adaptive Entropy-aware Optimization

链接：https://iclr.cc/virtual/2025/poster/28748 abstract： Test-time adaptation (TTA) has demonstrated significant potential in addressing distribution shifts between training and testing data. Open-set test-time adaptation (OSTTA) aims to adapt a source pre-trained model online to an unlabeled target domain that contains unknown classes. This task becomes more challenging when multiple modalities are involved. Existing methods have primarily focused on unimodal OSTTA, often filtering out low-confidence samples without addressing the complexities of multimodal data. In this work, we present Adaptive Entropy-aware Optimization (AEO), a novel framework specifically designed to tackle Multimodal Open-set Test-time Adaptation (MM-OSTTA) for the first time. Our analysis shows that the entropy difference between known and unknown samples in the target domain strongly correlates with MM-OSTTA performance. To leverage this, we propose two key components: Unknown-aware Adaptive Entropy Optimization (UAE) and Adaptive Modality Prediction Discrepancy Optimization (AMP). These components enhance the model’s ability to distinguish unknown class samples during online adaptation by amplifying the entropy difference between known and unknown samples. To thoroughly evaluate our proposed methods in the MM-OSTTA setting, we establish a new benchmark derived from existing datasets. This benchmark includes two downstream tasks – action recognition and 3D semantic segmentation – and incorporates five modalities: video, audio, and optical flow for action recognition, as well as LiDAR and camera for 3D semantic segmentation. Extensive experiments across various domain shift situations demonstrate the efficacy and versatility of the AEO framework. Additionally, we highlight the strong performance of AEO in long-term and continual MM-OSTTA settings, both of which are challenging and highly relevant to real-world applications. This underscores AEO’s robustness and adaptability in dynamic environments. Our source code and benchmarks are available at https://github.com/donghao51/AEO.

640. Local Loss Optimization in the Infinite Width: Stable Parameterization of Predictive Coding Networks and Target Propagation

链接：https://iclr.cc/virtual/2025/poster/28834 abstract： Local learning, which trains a network through layer-wise local targets and losses, has been studied as an alternative to backpropagation (BP) in neural computation. However, its algorithms often become more complex or require additional hyperparameters due to the locality, making it challenging to identify desirable settings where the algorithm progresses in a stable manner.To provide theoretical and quantitative insights, we introduce maximal update parameterization (μP) in the infinite-width limit for two representative designs of local targets: predictive coding (PC) and target propagation (TP). We verify that μP enables hyperparameter transfer across models of different widths.Furthermore, our analysis reveals unique and intriguing properties of μP that are not present in conventional BP. By analyzing deep linear networks, we find that PC's gradients interpolate between first-order and Gauss-Newton-like gradients, depending on the parameterization. We demonstrate that, in specific standard settings, PC in the infinite-width limit behaves more similarly to the first-order gradient.For TP, even with the standard scaling of the last layer differing from classical μP, its local loss optimization favors the feature learning regime over the kernel regime.

641. Provable Convergence and Limitations of Geometric Tempering for Langevin Dynamics

链接：https://iclr.cc/virtual/2025/poster/30453 abstract： Geometric tempering is a popular approach to sampling from challenging multi-modal probability distributions by instead sampling from a sequence of distributions which interpolate, using the geometric mean, between an easier proposal distribution and the target distribution. In this paper, we theoretically investigate the soundness of this approach when the sampling algorithm is Langevin dynamics, proving both upper and lower bounds. Our upper bounds are the first analysis in the literature under functional inequalities. They assert the convergence of tempered Langevin in continuous and discrete-time, and their minimization leads to closed-form optimal tempering schedules for some pairs of proposal and target distributions. Our lower bounds demonstrate a simple case where the geometric tempering takes exponential time, and further reveal that the geometric tempering can suffer from poor functional inequalities and slow convergence, even when the target distribution is well-conditioned. Overall, our results indicate that the geometric tempering may not help, and can even be harmful for convergence.

642. ShEPhERD: Diffusing shape, electrostatics, and pharmacophores for bioisosteric drug design

链接：https://iclr.cc/virtual/2025/poster/30057 abstract： Engineering molecules to exhibit precise 3D intermolecular interactions with their environment forms the basis of chemical design. In ligand-based drug design, bioisosteric analogues of known bioactive hits are often identified by virtually screening chemical libraries with shape, electrostatic, and pharmacophore

similarity scoring functions. We instead hypothesize that a generative model which learns the joint distribution over 3D molecular structures and their interaction profiles may facilitate 3D interaction-aware chemical design. We specifically design ShEPhERD, an SE(3)-equivariant diffusion model which jointly diffuses/denoises 3D molecular graphs and representations of their shapes, electrostatic potential surfaces, and (directional) pharmacophores to/from Gaussian noise. Inspired by traditional ligand discovery, we compose 3D similarity scoring functions to assess ShEPhERD's ability to conditionally generate novel molecules with desired interaction profiles. We demonstrate ShEPhERD's potential for impact via exemplary drug design tasks including natural product ligand hopping, protein-blind bioactive hit diversification, and bioisosteric fragment merging.

643. Simplifying Deep Temporal Difference Learning

链接：https://iclr.cc/virtual/2025/poster/30831 abstract： Q-learning played a foundational role in the field reinforcement learning (RL).However, TD algorithms with off-policy data, such as Q-learning, or nonlinear function approximation like deep neural networks require several additional tricks to stabilise training, primarily a large replay buffer and target networks. Unfortunately, the delayed updating of frozen network parameters in the target network harms the sample efficiency and, similarly, the large replay buffer introduces memory and implementation overheads. In this paper, we investigate whether it is possible to accelerate and simplify off-policy TD training while maintaining its stability. Our key theoretical result demonstrates for the first time that regularisation techniques such as LayerNorm can yield provably convergent TD algorithms without the need for a target network or replay buffer, even with off-policy data. Empirically, we find that online, parallelised sampling enabled by vectorised environments stabilises training without the need for a large replay buffer. Motivated by these findings, we propose PQN, our simplified deep online Q-Learning algorithm. Surprisingly, this simple algorithm is competitive with more complex methods like: Rainbow in Atari, PPO-RNN in Craftax, QMix in Smax, and can be up to 50x faster than traditional DQN without sacrificing sample efficiency. In an era where PPO has become the go-to RL algorithm, PQN reestablishes off-policy Q-learning as a viable alternative.

644. From Search to Sampling: Generative Models for Robust Algorithmic Recourse

链接：https://iclr.cc/virtual/2025/poster/29854 abstract： Algorithmic Recourse provides recommendations to individuals who are adversely impacted by automated model decisions, on how to alter their profiles to achieve a favorable outcome. Effective recourse methods must balance three conflicting goals: proximity to the original profile to minimize cost, plausibility for realistic recourse, and validity to ensure the desired outcome. We show that existing methods train for these objectives separately and then search for recourse through a joint optimization over the recourse goals during inference, leading to poor recourse recommendations. We introduce GenRe, a generative recourse model designed to train the three recourse objectives jointly. Training such generative models is non-trivial due to lack of direct recourse supervision. We propose efficient ways to synthesize such supervision and further show that GenRe's training leads to a consistent estimator. Unlike most prior methods, that employ non-robust gradient descent based search during inference, GenRe simply performs a forward sampling over the generative model to produce minimum cost recourse, leading to superior performance across multiple metrics. We also demonstrate GenRe provides the best trade-off between cost, plausibility and validity, compared to state-of-art baselines. We release anonymized code at: https://anonymous.4open.science/r/GenRe-BD71

645. Hot-pluggable Federated Learning: Bridging General and Personalized FL via Dynamic Selection

链接：https://iclr.cc/virtual/2025/poster/30591 abstract： Personalized federated learning (PFL) achieves high performance by assuming clients only meet test data locally, which does not meet many generic federated learning (GFL) scenarios. In this work, we theoretically show that PMs can be used to enhance GFL with a new learning problem named Selective FL (SFL), which involves optimizing PFL and model selection. However, storing and selecting whole models requires impractical computation and communication costs. To practically solve SFL, inspired by model components that attempt to edit a sub-model for specific purposes, we design an efficient and effective framework named Hot-Pluggable Federated Learning (HPFL). Specifically, clients individually train personalized plug-in modules based on a shared backbone, and upload them with a plug-in marker on the server modular store. In inference stage, an accurate selection algorithm allows clients to identify and retrieve suitable plug-in modules from the modular store to enhance their generalization performance on the target data distribution. Furthermore, we provide differential privacy protection during the selection with theoretical guarantee. Our comprehensive experiments and ablation studies demonstrate that HPFL significantly outperforms state-of-the-art GFL and PFL algorithms. Additionally, we empirically show HPFL's remarkable potential to resolve other practical FL problems such as continual federated learning and discuss its possible applications in one-shot FL, anarchic FL, and FL plug-in market. Our work is the first attempt towards improving GFL performance through a selecting mechanism with personalized plug-ins.

646. Machine Unlearning via Simulated Oracle Matching

链接：https://iclr.cc/virtual/2025/poster/31050 abstract： Machine unlearning---efficiently removing the effect of a small "forget set" of training data on a pre-trained machine learning model---has recently attracted significant research interest. Despite this interest, however, recent work shows that existing machine unlearning techniques do not hold up to thorough evaluation in non-convex settings. In this work, we introduce a new machine unlearning technique that exhibits strong empirical performance even in such challenging settings. Our starting point is the perspective that the goal of unlearning is to produce a model whose outputs

are statistically indistinguishable from those of a model re-trained on all but the forget set. This perspective naturally suggests a reduction from the unlearning problem to that of data attribution, where the goal is to predict the effect of changing the training set on a model's outputs. Thus motivated, we propose the following meta-algorithm, which we call Datamodel Matching (DMM): given a trained model, we (a) use data attribution to *predict the output of the model if it were re-trained on all but the forget set points; then (b) fine-tune the pre-trained model to match these predicted outputs. In a simple convex setting, we show how this approach provably outperforms a variety of iterative unlearning algorithms. Empirically, we use a combination of existing evaluations and a new metric based on the KL-divergence to show that even in non-convex settings, DMM achieves strong unlearning performance relative to existing algorithms. An added benefit of DMM is that it is a meta-algorithm, in the sense that future advances in data attribution translate directly into better unlearning algorithms, pointing to a clear direction for future progress in unlearning.

647. What Matters When Repurposing Diffusion Models for General Dense Perception Tasks?

链接：https://iclr.cc/virtual/2025/poster/30558 abstract： Extensive pre-training with large data is indispensable for downstream geometry and semantic visual perception tasks. Thanks to large-scale text-to-image (T2I) pretraining, recent works show promising results by simply fine-tuning T2I diffusion models for a few dense perception tasks. However, several crucial design decisions in this process still lack comprehensive justification, encompassing the necessity of the multi-step diffusion mechanism, training strategy, inference ensemble strategy, and fine-tuning data quality. In this work, we conduct a thorough investigation into critical factors that affect transfer efficiency and performance when using diffusion priors. Our key findings are: 1) High-quality fine-tuning data is paramount for both semantic and geometry perception tasks. 2) As a special case of the diffusion scheduler by setting its hyper-parameters, the multi-step generation can be simplified to a one-step fine-tuning paradigm without any loss of performance, while significantly speeding up inference. 3) Apart from fine-tuning the diffusion model with only latent space supervision, task-specific supervision can be beneficial to enhance fine-grained details. These observations culminate in the development of GenPercept, an effective deterministic one-step fine-tuning paradigm tailored for dense visual perception tasks exploiting diffusion priors. Different from the previous multi-step methods, our paradigm offers a much faster inference speed, and can be seamlessly integrated with customized perception decoders and loss functions for task-specific supervision, which can be critical for improving the fine-grained details of predictions. Comprehensive experiments on a diverse set of dense visual perceptual tasks, including monocular depth estimation, surface normal estimation, image segmentation, and matting, are performed to demonstrate the remarkable adaptability and effectiveness of our proposed method. Code: https://github.com/aim-uofa/GenPercept

648. Linear Partial Gromov-Wasserstein Embedding

链接：https://iclr.cc/virtual/2025/poster/30589 abstract： The Gromov–Wasserstein (GW) problem, a variant of the classical optimal transport (OT) problem, has attracted growing interest in the machine learning and data science communities due to its ability to quantify similarity between measures in different metric spaces. However, like the classical OT problem, GW imposes an equal mass constraint between measures, which restricts its application in many machine learning tasks. To address this limitation, the partial Gromov-Wasserstein (PGW) problem has been introduced. It relaxes the equal mass constraint, allowing the comparison of general positive Radon measures. Despite this, both GW and PGW face significant computational challenges due to their non-convex nature. To overcome these challenges, we propose the linear partial Gromov-Wasserstein (LPGW) embedding, a linearized embedding technique for the PGW problem. For K different metric measure spaces, the pairwise computation of the PGW distance requires solving the PGW problem $\mathcal{O}(K^2)$ times.In contrast, the proposed linearization technique reduces this to $\mathcal{O}(K)$ times. Similar to the linearization technique for the classical OT problem, we prove that LPGW defines a valid metric for metric measure spaces. Finally, we demonstrate the effectiveness of LPGW in practical applications such as shape retrieval and learning with transport-based embeddings, showing that LPGW preserves the advantages of PGW in partial matching while significantly enhancing computational efficiency. The code is available at https://github.com/mint-vu/Linearized_Partial_Gromov_Wasserstein.

649. Mitigate the Gap: Improving Cross-Modal Alignment in CLIP

链接：https://iclr.cc/virtual/2025/poster/29170 abstract： Contrastive Language--Image Pre-training (CLIP) has manifested remarkable improvements in zero-shot classification and cross-modal vision-language tasks. Yet, from a geometrical point of view, the CLIP embedding space has been found to have a pronounced modality gap. This gap renders the embedding space overly sparse and disconnected, with different modalities being densely distributed in distinct subregions of the hypersphere. In this work, we propose AlignCLIP, in order to improve the alignment between text and image embeddings, and thereby reduce the modality gap. AlignCLIP increases the cross-modal alignment, and yields gains across several zero-shot and fine-tuning downstream evaluations by sharing the learnable parameters between the modality encoders and a semantically-regularized separation objective function on the uni-modal embeddings. The source code and model checkpoints for reproducing our experiments are available at https://github.com/sarahESL/AlignCLIP.

650. Horizon Generalization in Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/30581 abstract： We study goal-conditioned RL through the lens of generalization, but not in the traditional sense of random augmentations and domain randomization. Rather, we aim to learn goal-directed policies that generalize with respect to the horizon: after training to reach nearby goals (which are easy to learn), these policies should

succeed in reaching distant goals (which are quite challenging to learn). In the same way that invariance is closely linked with generalization is other areas of machine learning (e.g., normalization layers make a network invariant to scale, and therefore generalize to inputs of varying scales), we show that this notion of horizon generalization is closely linked with invariance to planning: a policy navigating towards a goal will select the same actions as if it were navigating to a waypoint en route to that goal. Horizon generalization and invariance to planning are appealing because of their potential reach: they imply that a policy trained to reach nearby goals would succeed at reaching goals that are arbitrarily more distant.Our theoretical analysis proves that both horizon generalization and planning invariance are possible, under some assumptions. We present new experimental results, as well as recalling results from prior work, in support of our theoretical results. Taken together, our results open the door to studying how techniques for invariance and generalization developed in other areas of machine learning might be adapted to achieve this alluring property.

651. Robust Function-Calling for On-Device Language Model via Function Masking

链接：https://iclr.cc/virtual/2025/poster/27722 abstract： Large language models have demonstrated impressive value in performing as autonomous agents when equipped with external tools and API calls. Nonetheless, effectively harnessing their potential for executing complex tasks crucially relies on enhancements in their function-calling capabilities. This paper identifies a critical gap in existing function-calling models, where performance varies significantly across benchmarks, often due to over-fitting to specific naming conventions. To address such an issue, we introduce Hammer, a novel family of foundation models specifically engineered for on-device function calling. Hammer employs an augmented dataset that enhances models' sensitivity to irrelevant functions and incorporates function masking techniques to minimize over-fitting. Our empirical evaluations reveal that Hammer not only outperforms larger models but also demonstrates robust generalization across diverse benchmarks, achieving state-of-the-art results. Our open-source contributions include a specialized dataset for irrelevance detection, a tuning framework for enhanced generalization, and the Hammer models, establishing a new standard for function-calling performance.

652. Improved Sampling Of Diffusion Models In Fluid Dynamics With Tweedie's Formula

链接：https://iclr.cc/virtual/2025/poster/31263 abstract： State-of-the-art Denoising Diffusion Probabilistic Models (DDPMs) rely on an expensive sampling process with a large Number of Function Evaluations (NFEs) to provide high-fidelity predictions. This computational bottleneck renders diffusion models less appealing as surrogates for the spatio-temporal prediction of physics-based problems with long rollout horizons. We propose Truncated Sampling Models, enabling single-step and few-step sampling with elevated fidelity by simple truncation of the diffusion process, reducing the gap between DDPMs and deterministic single-step approaches. We also introduce a novel approach, Iterative Refinement, to sample pre-trained DDPMs by reformulating the generative process as a refinement process with few sampling steps. Both proposed methods enable significant improvements in accuracy compared to DDPMs, DDIMs, and EDMs with NFEs $\leq$ 10 on a diverse set of experiments, including incompressible and compressible turbulent flow and airfoil flow uncertainty simulations. Our proposed methods provide stable predictions for long rollout horizons in time-dependent problems and are able to learn all modes of the data distribution in steady-state problems with high uncertainty.

653. Temporal Flexibility in Spiking Neural Networks: Towards Generalization Across Time Steps and Deployment Friendliness

链接：https://iclr.cc/virtual/2025/poster/30705 abstract： Spiking Neural Networks (SNNs), models inspired by neural mechanisms in the brain, allow for energy-efficient implementation on neuromorphic hardware. However, SNNs trained with current direct training approaches are constrained to a specific time step. This "temporal inflexibility" 1) hinders SNNs' deployment on time-step-free fully event-driven chips and 2) prevents energy-performance balance based on dynamic inference time steps. In this study, we first explore the feasibility of training SNNs that generalize across different time steps. We then introduce Mixed Time-step Training (MTT), a novel method that improves the temporal flexibility of SNNs, making SNNs adaptive to diverse temporal structures. During each iteration of MTT, random time steps are assigned to different SNN stages, with spikes transmitted between stages via communication modules. After training, the weights are deployed and evaluated on both time-stepped and fully event-driven platforms. Experimental results show that models trained by MTT gain remarkable temporal flexibility, friendliness for both event-driven and clock-driven deployment (nearly lossless on N-MNIST and 10.1\% higher than standard methods on CIFAR10-DVS), enhanced network generalization, and near SOTA performance. To the best of our knowledge, this is the first work to report the results of large-scale SNN deployment on fully event-driven scenarios.

654. SeCom: On Memory Construction and Retrieval for Personalized Conversational Agents

链接：https://iclr.cc/virtual/2025/poster/27790 abstract： To deliver coherent and personalized experiences in long-term conversations, existing approaches typically perform retrieval augmented response generation by constructing memory banks from conversation history at either the turn-level, session-level, or through summarization techniques.In this paper, we explore the impact of different memory granularities and present two key findings: (1) Both turn-level and session-level memory units are suboptimal, affecting not only the quality of final responses, but also the accuracy of the retrieval process.(2) The redundancy in

natural language introduces noise, hindering precise retrieval. We demonstrate that LLMLingua-2, originally designed for prompt compression to accelerate LLM inference, can serve as an effective denoising method to enhance memory retrieval accuracy.Building on these insights, we propose SeCom, a method that constructs a memory bank with topical segments by introducing a conversation Segmentation model, while performing memory retrieval based on Compressed memory units.Experimental results show that SeCom outperforms turn-level, session-level, and several summarization-based methods on long-term conversation benchmarks such as LOCOMO and Long-MT-Bench+. Additionally, the proposed conversation segmentation method demonstrates superior performance on dialogue segmentation datasets such as DialSeg711, TIAGE, and SuperDialSeg.

655. Determine-Then-Ensemble: Necessity of Top-k Union for Large Language Model Ensembling

链接：https://iclr.cc/virtual/2025/poster/30354 abstract： Large language models (LLMs) exhibit varying strengths and weaknesses across different tasks, prompting recent studies to explore the benefits of ensembling models to leverage their complementary advantages. However, existing LLM ensembling methods often overlook model compatibility and struggle with inefficient alignment of probabilities across the entire vocabulary. In this study, we empirically investigate the factors influencing ensemble performance, identifying model performance, vocabulary size, and response style as key determinants, revealing that compatibility among models is essential for effective ensembling. This analysis leads to the development of a simple yet effective model selection strategy that identifies compatible models. Additionally, we introduce the \textsc{Uni}on \textsc{T}op-k \textsc{E}nsembling (\textsc{UniTE}), a novel approach that efficiently combines models by focusing on the union of the top-k tokens from each model, thereby avoiding the need for full vocabulary alignment and reducing computational overhead. Extensive evaluations across multiple benchmarks demonstrate that \textsc{UniTE} significantly enhances performance compared to existing methods, offering a more efficient framework for LLM ensembling.

656. Finding Shared Decodable Concepts and their Negations in the Brain

链接：https://iclr.cc/virtual/2025/poster/30022 abstract： Prior work has offered evidence for functional localization in the brain; different anatomical regions preferentially activate for certain types of visual input. For example, the fusiform face area preferentially activates for visual stimuli that include a face. However, the spectrum of visual semantics is extensive, and only a few semantically-tuned patches of cortex have so far been identified in the human brain. Using a multimodal (natural language and image) neural network architecture (CLIP, \cite{CLIP}, we train a highly accurate contrastive model that maps brain responses during naturalistic image viewing to CLIP embeddings. We then use a novel adaptation of the DBSCAN clustering algorithm to cluster the parameters of these participant-specific contrastive models. This reveals what we call Shared Decodable Concepts (SDCs): clusters in CLIP space that are decodable from common sets of voxels across multiple participants.Examining the images most and least associated with each SDC cluster gives us additional insight into the semantic properties of each SDC. We note SDCs for previously reported visual features (e.g. orientation tuning in early visual cortex) as well as visual semantic concepts such as faces, places and bodies. In cases where our method finds multiple clusters for a visuo-semantic concept, the least associated images allow us to dissociate between confounding factors. For example, we discovered two clusters of food images, one driven by color, the other by shape. We also uncover previously unreported areas with visuo-semantic sensitivity such as regions of extrastriate body area (EBA) tuned for legs/hands and sensitivity to numerosity in right intraparietal sulcus, sensitivity associated with visual perspective (close/far) and more. Thus, our contrastive-learning methodology better characterizes new and existing visuo-semantic representations in the brain by leveraging multimodal neural network representations and a novel adaptation of clustering algorithms.

657. State Space Model Meets Transformer: A New Paradigm for 3D Object Detection

链接：https://iclr.cc/virtual/2025/poster/29524 abstract： DETR-based methods, which use multi-layer transformer decoders to refine object queries iteratively, have shown promising performance in 3D indoor object detection. However, the scene point features in the transformer decoder remain fixed, leading to minimal contributions from later decoder layers, thereby limiting performance improvement. Recently, State Space Models (SSM) have shown efficient context modeling ability with linear complexity through iterative interactions between system states and inputs. Inspired by SSMs, we propose a new 3D object DEtection paradigm with an interactive STate space model (DEST). In the interactive SSM, we design a novel state-dependent SSM parameterization method that enables system states to effectively serve as queries in 3D indoor detection tasks. In addition, we introduce four key designs tailored to the characteristics of point cloud and SSM: The serialization and bidirectional scanning strategies enable bidirectional feature interaction among scene points within the SSM. The inter-state attention mechanism models the relationships between state points, while the gated feed-forward network enhances inter-channel correlations. To the best of our knowledge, this is the first method to model queries as system states and scene points as system inputs, which can simultaneously update scene point features and query features with linear complexity. Extensive experiments on two challenging datasets demonstrate the effectiveness of our DEST-based method. Our method improves the GroupFree baseline in terms of AP_{50} on ScanNet V2 (+5.3) and SUN RGB-D (+3.2) datasets. Based on the VDETR baseline, Our method sets a new state-of-the-art on the ScanNetV2 and SUN RGB-D datasets.

658. Scaling Optimal LR Across Token Horizons

链接：https://iclr.cc/virtual/2025/poster/29358 abstract： State-of-the-art LLMs are powered by scaling -- scaling model size, training tokens, and cluster size. It is economically infeasible to extensively tune hyperparameters for the largest runs. Instead, approximately optimal hyperparameters must be inferred or transferred from smaller experiments. Hyperparameter transfer across model sizes has been studied in muP. However, hyperparameter transfer across training tokens -- or token horizon -- has not been studied yet. To remedy this we conduct a large-scale empirical study on how optimal learning rate (LR) depends on the token horizon in LLM training. We first demonstrate that the optimal LR changes significantly with token horizon -- longer training necessitates smaller LR. Secondly, we demonstrate that the optimal LR follows a scaling law and that the optimal LR for longer horizons can be accurately estimated from shorter horizons via such scaling laws. We also provide a rule-of-thumb for transferring LR across token horizons with zero overhead over current practices. Lastly, we provide evidence that LLama-1 used too high LR, and thus argue that hyperparameter transfer across data size is an overlooked component of LLM training.

659. A Probabilistic Perspective on Unlearning and Alignment for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30975 abstract： Comprehensive evaluation of Large Language Models (LLMs) is an open research problem. Existing evaluations rely on deterministic point estimates generated via greedy decoding. However, we find that deterministic evaluations fail to capture the whole output distribution of a model, yielding inaccurate estimations of model capabilities. This is particularly problematic in critical contexts such as unlearning and alignment, where precise model evaluations are crucial. To remedy this, we introduce the first formal probabilistic evaluation framework for LLMs. Namely, we propose novel metrics with high probability guarantees concerning the output distribution of a model. Our metrics are application-independent and allow practitioners to make more reliable estimates about model capabilities before deployment. Our experimental analysis reveals that deterministic evaluations falsely indicate successful unlearning and alignment, whereas our probabilistic evaluations better capture model capabilities. We show how to overcome challenges associated with probabilistic outputs in a case study on unlearning by introducing (1) a novel loss based on entropy optimization, and (2) adaptive temperature scaling. We demonstrate that our approach significantly enhances unlearning in probabilistic settings on recent benchmarks. Overall, our proposed shift from point estimates to probabilistic evaluations of output distributions represents an important step toward comprehensive evaluations of LLMs.

660. Flow Matching with Gaussian Process Priors for Probabilistic Time Series Forecasting

链接：https://iclr.cc/virtual/2025/poster/27944 abstract： Recent advancements in generative modeling, particularly diffusion models, have opened new directions for time series modeling, achieving state-of-the-art performance in forecasting and synthesis. However, the reliance of diffusion-based models on a simple, fixed prior complicates the generative process since the data and prior distributions differ significantly. We introduce TSFlow, a conditional flow matching (CFM) model for time series combining Gaussian processes, optimal transport paths, and data-dependent prior distributions. By incorporating (conditional) Gaussian processes, TSFlow aligns the prior distribution more closely with the temporal structure of the data, enhancing both unconditional and conditional generation. Furthermore, we propose conditional prior sampling to enable probabilistic forecasting with an unconditionally trained model. In our experimental evaluation on eight real-world datasets, we demonstrate the generative capabilities of TSFlow, producing high-quality unconditional samples. Finally, we show that both conditionally and unconditionally trained models achieve competitive results across multiple forecasting benchmarks.

661. A primer on analytical learning dynamics of nonlinear neural networks

链接：https://iclr.cc/virtual/2025/poster/31338 abstract： The learning dynamics of neural networks—in particular, how parameters change over time during training—describe how data, architecture, and algorithm interact in time to produce a trained neural network model. Characterizing these dynamics, in general, remains an open problem in machine learning, but, handily, restricting the setting allows careful empirical studies and even analytical results. In this blog post, we review approaches to analyzing the learning dynamics of nonlinear neural networks, focusing on a particular setting known as teacher-student that permits an explicit analytical expression for the generalization error of a nonlinear neural network trained with online gradient descent. We provide an accessible mathematical formulation of this analysis and a JAX codebase to implement simulation of the analytical system of ordinary differential equations alongside neural network training in this setting. We conclude with a discussion of how this analytical paradigm has been used to investigate generalization in neural networks and beyond.

662. Bio-xLSTM: Generative modeling, representation and in-context learning of biological and chemical sequences

链接：https://iclr.cc/virtual/2025/poster/30154 abstract： Language models for biological and chemical sequences enable crucial applications such as drug discovery, protein engineering, and precision medicine. Currently, these language models are predominantly based on Transformer architectures. While Transformers have yielded impressive results, their quadratic runtime dependency on sequence length complicates their use for long genomic sequences and in-context learning on proteins and chemical sequences. Recently, the recurrent xLSTM architecture has been shown to perform favorably compared to Transformers and modern state-space models (SSMs) in the natural language domain. Similar to SSMs, xLSTMs have linear

runtime dependency and allow for constant-memory decoding at inference time, which makes them prime candidates for modeling long-range dependencies in biological and chemical sequences. In this work, we tailor xLSTM towards these domains and we propose a suite of language models called Bio-xLSTM. Extensive experiments in three large domains, genomics, proteins, and chemistry, were performed to assess xLSTM's ability to model biological and chemical sequences. The results show that Bio-xLSTM is a highly proficient generative model for DNA, protein, and chemical sequences, learns rich representations, and can perform in-context learning for proteins and small molecules.

663. Enhancing Clustered Federated Learning: Integration of Strategies and Improved Methodologies

链接：https://iclr.cc/virtual/2025/poster/27669 abstract： Federated Learning (FL) is an evolving distributed machine learning approach that safeguards client privacy by keeping data on edge devices. However, the variation in data among clients poses challenges in training models that excel across all local distributions. Recent studies suggest clustering as a solution to address client heterogeneity in FL by grouping clients with distribution shifts into distinct clusters. Nonetheless, the diverse learning frameworks used in current clustered FL methods create difficulties in integrating these methods, leveraging their advantages, and making further enhancements. To this end, this paper conducts a thorough examination of existing clustered FL methods and introduces a four-tier framework, named HCFL, to encompass and extend the existing approaches. Utilizing the HCFL, we identify persistent challenges associated with current clustering methods in each tier and propose an enhanced clustering method called HCFL^{+} to overcome these challenges. Through extensive numerical evaluations, we demonstrate the effectiveness of our clustering framework and the enhanced components. Our code is available at \url{https://github.com/LINs-lab/HCFL}.

664. KiVA: Kid-inspired Visual Analogies for Testing Large Multimodal Models

链接：https://iclr.cc/virtual/2025/poster/27918 abstract： This paper investigates visual analogical reasoning in large multimodal models (LMMs) compared to human adults and children. A "visual analogy" is an abstract rule inferred from one image and applied to another. While benchmarks exist for testing visual reasoning in LMMs, they require advanced skills and omit basic visual analogies that even young children can make. Inspired by developmental psychology, we propose a new benchmark of 4,300 visual transformations of everyday objects to test LMMs on visual analogical reasoning and compare them to children (ages three to five) and to adults. We structure the evaluation into three stages: identifying what changed (e.g., color, number, etc.), how it changed (e.g., added one object), and applying the rule to new scenarios. Our findings show that while GPT-o1, GPT-4V, LLaVA-1.5, and MANTIS identify the "what" effectively, they struggle with quantifying the "how" and extrapolating this rule to new objects. In contrast, children and adults exhibit much stronger analogical reasoning at all three stages. Additionally, the strongest tested model, GPT-o1, performs better in tasks involving simple surface-level visual attributes like color and size, correlating with quicker human adult response times. Conversely, more complex tasks such as number, rotation, and reflection, which necessitate extensive cognitive processing and understanding of extrinsic spatial properties in the physical world, present more significant challenges. Altogether, these findings highlight the limitations of training models on data that primarily consists of 2D images and text.

665. Can LLMs Understand Time Series Anomalies?

链接：https://iclr.cc/virtual/2025/poster/30008 abstract： Large Language Models (LLMs) have gained popularity in time series forecasting, but their potential for anomaly detection remains largely unexplored. Our study investigates whether LLMs can understand and detect anomalies in time series data, focusing on zero-shot and few-shot scenarios. Inspired by conjectures about LLMs' behavior from time series forecasting research, we formulate key hypotheses about LLMs' capabilities in time series anomaly detection. We design and conduct principled experiments to test each of these hypotheses. Our investigation reveals several surprising findings about LLMs for time series: (1) LLMs understand time series better as images rather than as text, (2) LLMs do not demonstrate enhanced performance when prompted to engage in explicit reasoning about time series analysis. (3) Contrary to common beliefs, LLMs' understanding of time series do not stem from their repetition biases or arithmetic abilities. (4) LLMs' behaviors and performance in time series analysis vary significantly across different models. This study provides the first comprehensive analysis of contemporary LLM capabilities in time series anomaly detection. Our results suggest that while LLMs can understand trivial time series anomalies (we have no evidence that they can understand more subtle real-world anomalies), many common conjectures based on their reasoning capabilities do not hold. All synthetic dataset generators, final prompts, and evaluation scripts have been made available in https://github.com/rose-stl-lab/anomllm.

666. TRACE: Temporal Grounding Video LLM via Causal Event Modeling

链接：https://iclr.cc/virtual/2025/poster/31224 abstract： Video Temporal Grounding (VTG) is a crucial capability for video understanding models and plays a vital role in downstream tasks such as video browsing and editing. To effectively handle various tasks simultaneously and enable zero-shot prediction, there is a growing trend in employing video LLMs for VTG tasks. However, current video LLM-based methods rely exclusively on natural language generation, lacking the ability to model the clear structure inherent in videos, which restricts their effectiveness in tackling VTG tasks. To address this issue, this paper first formally introduces causal event modeling framework, which represents video LLM outputs as sequences of events, and predict the current event using previous events, video inputs, and textural instructions. Each event consists of three components:

timestamps, salient scores, and textual captions. We then propose a novel task-interleaved video LLM called TRACE to effectively implement the causal event modeling framework in practice. The TRACE process visual frames, timestamps, salient scores, and text as distinct tasks, employing various encoders and decoding heads for each. Task tokens are arranged in an interleaved sequence according to the causal event modeling framework's formulation.Extensive experiments on various VTG tasks and datasets demonstrate the superior performance of TRACE compared to state-of-the-art video LLMs. Our model and code are avaliable at \url{https://github.com/gyxxyg/TRACE}.

667. Timer-XL: Long-Context Transformers for Unified Time Series Forecasting

链接：https://iclr.cc/virtual/2025/poster/30062 abstract： We present Timer-XL, a causal Transformer for unified time series forecasting. To uniformly predict multidimensional time series, we generalize next token prediction, predominantly adopted for 1D token sequences, to multivariate next token prediction. The paradigm formulates various forecasting tasks as a long-context prediction problem. We opt for decoder-only Transformers that capture causal dependencies from varying-length contexts for unified forecasting, making predictions on non-stationary univariate time series, multivariate series with complicated dynamics and correlations, as well as covariate-informed contexts that include exogenous variables. Technically, we propose a universal TimeAttention to capture fine-grained intra- and inter-series dependencies of flattened time series tokens (patches), which is further enhanced by deft position embedding for temporal causality and variable equivalence. Timer-XL achieves state-of-the-art performance across task-specific forecasting benchmarks through a unified approach. Based on large-scale pre-training, Timer-XL achieves state-of-the-art zero-shot performance, making it a promising architecture for pre-trained time series models. Code is available at this repository: https://github.com/thuml/Timer-XL.

668. ClimaQA: An Automated Evaluation Framework for Climate Question Answering Models

链接：https://iclr.cc/virtual/2025/poster/28801 abstract： The use of Large Language Models (LLMs) in climate science has recently gained significant attention. However, a critical issue remains: the lack of a comprehensive evaluation framework capable of assessing the quality and scientific validity of model outputs. To address this issue, we develop ClimaGen (Climate QA Generator), an adaptive learning framework that generates question-answer pairs from graduate textbooks with climate scientists in the loop. As a result, we present ClimaQA-Gold, an expert-annotated benchmark dataset alongside ClimaQA-Silver, a large-scale, comprehensive synthetic QA dataset for climate science. Finally, we develop evaluation strategies and compare different LLMs on our benchmarks. Our results offer novel insights into various approaches used to enhance knowledge of climate LLMs. ClimaQA’s source code is publicly available at https://github.com/Rose-STL-Lab/genie-climaqa

669. Neural Fluid Simulation on Geometric Surfaces

链接：https://iclr.cc/virtual/2025/poster/30967 abstract： Incompressible fluid on the surface is an interesting research area in the fluid simulation, which is the fundamental building block in visual effects, design of liquid crystal films, scientific analyses of atmospheric and oceanic phenomena, etc. The task brings two key challenges: the extension of the physical laws on 3D surfaces and the preservation of the energy and volume. Traditional methods rely on grids or meshes for spatial discretization, which leads to high memory consumption and a lack of robustness and adaptivity for various mesh qualities and representations. Many implicit representations based simulators like INSR are proposed for the storage efficiency and continuity, but they face challenges in the surface simulation and the energy dissipation. We propose a neural physical simulation framework on the surface with the implicit neural representation. Our method constructs a parameterized vector field with the exterior calculus and Closest Point Method on the surfaces, which guarantees the divergence-free property and enables the simulation on different surface representations (e.g. implicit neural represented surfaces). We further adopt a corresponding covariant derivative based advection process for surface flow dynamics and energy preservation. Our method shows higher accuracy, flexibility and memory-efficiency in the simulations of various surfaces with low energy dissipation. Numerical studies also highlight the potential of our framework across different practical applications such as vorticity shape generation and vector field Helmholtz decomposition.

670. Recovering Manifold Structure Using Ollivier Ricci Curvature

链接：https://iclr.cc/virtual/2025/poster/29161 abstract： We introduce ORC-ManL, a new algorithm to prune spurious edges from nearest neighbor graphs using a criterion based on Ollivier-Ricci curvature and estimated metric distortion. Our motivation comes from manifold learning: we show that when the data generating the nearest-neighbor graph consists of noisy samples from a low-dimensional manifold, edges that shortcut through the ambient space have more negative Ollivier-Ricci curvature than edges that lie along the data manifold. We demonstrate that our method outperforms alternative pruning methods and that it significantly improves performance on many downstream geometric data analysis tasks that use nearest neighbor graphs as input. Specifically, we evaluate on manifold learning, persistent homology, dimension estimation, and others. We also show that ORC-ManL can be used to improve clustering and manifold learning of single-cell RNA sequencing data. Finally, we provide empirical convergence experiments that support our theoretical findings.

671. Investigating the Pre-Training Dynamics of In-Context Learning: Task

Recognition vs. Task Learning

链接：https://iclr.cc/virtual/2025/poster/28736 abstract： The emergence of in-context learning (ICL) is potentially attributed to two major abilities: task recognition (TR) for recognizing the task from demonstrations and utilizing pre-trained priors, and task learning (TL) for learning from demonstrations. However, relationships between the two abilities and how such relationships affect the emergence of ICL is unclear. In this paper, we take the first step by examining the pre-training dynamics of the emergence of ICL. With carefully designed metrics, we find that these two abilities are, in fact, competitive during pre-training. Moreover, we observe a negative correlation between the competition and the performance of ICL. Further analysis of common pre-training factors (i.e., model size, dataset size, and data curriculum) demonstrates possible ways to regulate the competition. Based on these insights, we propose a simple yet effective method to better integrate these two abilities for ICL at inference time. Through adaptive ensemble learning, the performance of ICL can be significantly boosted, enabling two small models to outperform a larger one with more than twice the parameters.

672. Towards Self-Supervised Covariance Estimation in Deep Heteroscedastic Regression

链接：https://iclr.cc/virtual/2025/poster/29712 abstract： Deep heteroscedastic regression models the mean and covariance of the target distribution through neural networks. The challenge arises from heteroscedasticity, which implies that the covariance is sample dependent and is often unknown. Consequently, recent methods learn the covariance through unsupervised frameworks, which unfortunately yield a trade-off between computational complexity and accuracy. While this trade-off could be alleviated through supervision, obtaining labels for the covariance is non-trivial.Here, we study self-supervised covariance estimation in deep heteroscedastic regression. We address two questions: (1) How should we supervise the covariance assuming ground truth is available? (2) How can we obtain pseudo labels in the absence of the ground-truth? We address (1) by analysing two popular measures: the KL Divergence and the 2-Wasserstein distance. Subsequently, we derive an upper bound on the 2-Wasserstein distance between normal distributions with non-commutative covariances that is stable to optimize. We address (2) through a simple neighborhood based heuristic algorithm which results in surprisingly effective pseudo labels for the covariance. Our experiments over a wide range of synthetic and real datasets demonstrate that the proposed 2-Wasserstein bound coupled with pseudo label annotations results in a computationally cheaper yet accurate deep heteroscedastic regression.

673. The Complexity of Two-Team Polymatrix Games with Independent Adversaries

链接：https://iclr.cc/virtual/2025/poster/30687 abstract： Adversarial multiplayer games are an important object of study in multiagent learning. In particular, polymatrix zero-sum games are a multiplayer setting where Nash equilibria are known to be efficiently computable. Towards understanding the limits of tractability in polymatrix games, we study the computation of Nash equilibria in such games where each pair of players plays either a zero-sum or a coordination game. We are particularly interested in the setting where players can be grouped into a small number of teams of identical interest. While the three-team version of the problem is known to be PPAD-complete, the complexity for two teams has remained open. Our main contribution is to prove that the two-team version remains hard, namely it is CLS-hard. Furthermore, we show that this lower bound is tight for the setting where one of the teams consists of multiple independent adversaries. On the way to obtaining our main result, we prove hardness of finding any stationary point in the simplest type of non-convex-concave min-max constrained optimization problem, namely for a class of bilinear polynomial objective functions.

674. EgoSim: Egocentric Exploration in Virtual Worlds with Multi-modal Conditioning

链接：https://iclr.cc/virtual/2025/poster/27687 abstract： Recent advancements in video diffusion models have established a strong foundation for developing world models with practical applications. The next challenge lies in exploring how an agent can leverage these foundation models to understand, interact with, and plan within observed environments. This requires adding more controllability to the model, transforming it into a versatile game engine capable of dynamic manipulation and control. To address this, we investigated three key conditioning factors: camera, context frame, and text, identifying limitations in current model designs. Specifically, the fusion of camera embeddings with video features leads to camera control being influenced by those features. Additionally, while textual information compensates for necessary spatiotemporal structures, it often intrudes into already observed parts of the scene. To tackle these issues, we designed the Spacetime Epipolar Attention Layer, which ensures that egomotion generated by the model strictly aligns with the camera’s movement through rigid constraints. Moreover, we propose the CI2V-adapter, which uses camera information to better determine whether to prioritize textual or visual embeddings, thereby alleviating the issue of textual intrusion into observed areas. Through extensive experiments, we demonstrate that our new model EgoSim achieves excellent results on both the RealEstate and newly repurposed Epic-Field datasets. For more results, please refer to https://egosim.github.io/EgoSim/.

675. PWM: Policy Learning with Multi-Task World Models

链接：https://iclr.cc/virtual/2025/poster/28766 abstract： Reinforcement Learning (RL) has made significant strides in complex tasks but struggles in multi-task settings with different embodiments. World model methods offer scalability by learning a

simulation of the environment but often rely on inefficient gradient-free optimization methods for policy extraction. In contrast, gradient-based methods exhibit lower variance but fail to handle discontinuities. Our work reveals that well-regularized world models can generate smoother optimization landscapes than the actual dynamics, facilitating more effective first-order optimization. We introduce Policy learning with multi-task World Models (PWM), a novel model-based RL algorithm for continuous control. Initially, the world model is pre-trained on offline data, and then policies are extracted from it using first-order optimization in less than 10 minutes per task. PWM effectively solves tasks with up to 152 action dimensions and outperforms methods that use ground-truth dynamics. Additionally, PWM scales to an 80-task setting, achieving up to 27\% higher rewards than existing baselines without relying on costly online planning. Visualizations and code are available at imgeorgiev.com/pwm.

676. MRAG-Bench: Vision-Centric Evaluation for Retrieval-Augmented Multimodal Models

链接：https://iclr.cc/virtual/2025/poster/29447 abstract： Existing multimodal retrieval benchmarks primarily focus on evaluating whether models can retrieve and utilize external textual knowledge for question answering. However, there are scenarios where retrieving visual information is either more beneficial or easier to access than textual data. In this paper, we introduce a multimodal retrieval-augmented generation benchmark, MRAG-Bench, in which we systematically identify and categorize scenarios where visually augmented knowledge is better than textual knowledge, for instance, more images from varying viewpoints.MRAG-Bench consists of 16,130 images and 1,353 human-annotated multiple-choice questions across 9 distinct scenarios. With MRAG-Bench, we conduct an evaluation of 10 open-source and 4 proprietary large vision-language models (LVLMs). Our results show that all LVLMs exhibit greater improvements when augmented with images compared to textual knowledge, confirming that MRAG-Bench is vision-centric. Additionally, we conduct extensive analysis with MRAG-Bench, which offers valuable insights into retrieval-augmented LVLMs. Notably, the top-performing model, GPT-4o, faces challenges in effectively leveraging retrieved knowledge, achieving only a 5.82\% improvement with ground-truth information, in contrast to a 33.16\% improvement observed in human participants. These findings highlight the importance of MRAG-Bench in encouraging the community to enhance LVLMs' ability to utilize retrieved visual knowledge more effectively.

677. Distilling Reinforcement Learning Algorithms for In-Context Model-Based Planning

链接：https://iclr.cc/virtual/2025/poster/30559 abstract： Recent studies have shown that Transformers can perform in-context reinforcement learning (RL) by imitating existing RL algorithms, enabling sample-efficient adaptation to unseen tasks without parameter updates. However, these models also inherit the suboptimal behaviors of the RL algorithms they imitate. This issue primarily arises due to the gradual update rule employed by those algorithms. Model-based planning offers a promising solution to this limitation by allowing the models to simulate potential outcomes before taking action, providing an additional mechanism to deviate from the suboptimal behavior. Rather than learning a separate dynamics model, we propose Distillation for In-Context Planning (DICP), an in-context model-based RL framework where Transformers simultaneously learn environment dynamics and improve policy in-context. We evaluate DICP across a range of discrete and continuous environments, including Darkroom variants and Meta-World. Our results show that DICP achieves state-of-the-art performance while requiring significantly fewer environment interactions than baselines, which include both model-free counterparts and existing meta-RL methods.

678. Breaking the $\log(1/\Delta_2)$ Barrier: Better Batched Best Arm Identification with Adaptive Grids

链接：https://iclr.cc/virtual/2025/poster/29082 abstract： We investigate the problem of batched best arm identification in multi-armed bandits, where we want to find the best arm from a set of n arms while minimizing both the number of samples and batches. We introduce an algorithm that achieves near-optimal sample complexity and features an instance-sensitive batch complexity, which breaks the $\log(1/\Delta_2)$ barrier. The main contribution of our algorithm is a novel sample allocation scheme that effectively balances exploration and exploitation for batch sizes. Experimental results indicate that our approach is more batch-efficient across various setups. We also extend this framework to the problem of batched best arm identification in linear bandits and achieve similar improvements.

679. Learning from Imperfect Human Feedback: A Tale from Corruption-Robust Dueling

链接：https://iclr.cc/virtual/2025/poster/28276 abstract： This paper studies Learning from Imperfect Human Feedback (LIHF), addressing the potential irrationality or imperfect perception when learning from comparative human feedback. Building on evidences that human's imperfection decays over time (i.e., humans learn to improve), we cast this problem as a concave-utility continuous-action dueling bandit but under a restricted form of corruption: i.e., the corruption scale is decaying over time as $t^{\rho-1}$ for some ``imperfection rate" $\rho \in [0, 1]$. With T as the total number of iterations, we establish a regret lower bound of $ \Omega(\max{\sqrt{T}, T^{\rho}})$ for LIHF, even when ρ is known. For the same setting, we develop the Robustified Stochastic Mirror Descent for Imperfect Dueling (RoSMID) algorithm, which achieves nearly optimal regret $\tilde{\mathcal{O}}(\max{\sqrt{T}, T^{\rho}})$. Core to our analysis is a novel framework for analyzing gradient-based algorithms for dueling bandit under corruption, and we demonstrate its general applicability by showing how this framework can be easily applied to obtain corruption-robust guarantees for other popular gradient-based dueling bandit algorithms. Our theoretical

results are validated by extensive experiments.

680. On the Expressiveness of Rational ReLU Neural Networks With Bounded Depth

链接：https://iclr.cc/virtual/2025/poster/27977 abstract： To confirm that the expressive power of ReLU neural networks grows with their depth, the function $F_n = \max (0,x_1,\ldots,x_n)$ has been considered in the literature. A conjecture by Hertrich, Basu, Di Summa, and Skutella [NeurIPS 2021] states that any ReLU network that exactly represents F_n has at least $\lceil \log_2 (n+1) \rceil$ hidden layers. The conjecture has recently been confirmed for networks with integer weights by Haase, Hertrich, and Loho [ICLR 2023]. We follow up on this line of research and show that, within ReLU networks whose weights are decimal fractions, F_n can only be represented by networks with at least $\lceil \log_3 (n+1) \rceil$ hidden layers. Moreover, if all weights are N-ary fractions, then F_n can only be represented by networks with at least $\Omega(\frac{\ln n}{\ln \ln N})$ layers. These results are a partial confirmation of the above conjecture for rational ReLU networks, and provide the first non-constant lower bound on the depth of practically relevant ReLU networks.

681. Transformer Learns Optimal Variable Selection in Group-Sparse Classification

链接：https://iclr.cc/virtual/2025/poster/28844 abstract： Transformers have demonstrated remarkable success across various applications. However, the success of transformers have not been understood in theory. In this work, we give a case study of how transformers can be trained to learn a classic statistical model with "group sparsity", where the input variables form multiple groups, and the label only depends on the variables from one of the groups. We theoretically demonstrate that, a one-layer transformer trained by gradient descent can correctly leverage the attention mechanism to select variables, disregarding irrelevant ones and focusing on those beneficial for classification. We also demonstrate that a well-pretrained one-layer transformer can be adapted to new downstream tasks to achieve good prediction accuracy with a limited number of samples. Our study sheds light on how transformers effectively learn structured data.

682. Graph Neural Networks Gone Hogwild

链接：https://iclr.cc/virtual/2025/poster/29350 abstract： Graph neural networks (GNNs) appear to be powerful tools to learn state representations for agents in distributed, decentralized multi-agent systems, but generate catastrophically incorrect predictions when nodes update asynchronously during inference. This failure under asynchrony effectively excludes these architectures from many potential applications where synchrony is difficult or impossible to enforce, e.g., robotic swarms or sensor networks. In this work we identify "implicitly-defined" GNNs as a class of architectures which is provably robust to asynchronous "hogwild" inference, adapting convergence guarantees from work in asynchronous and distributed optimization. We then propose a novel implicitly-defined GNN architecture, which we call an energy GNN. We show that this architecture outperforms other GNNs from this class on a variety of synthetic tasks inspired by multi-agent systems.

683. Adaptive Q-Network: On-the-fly Target Selection for Deep Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28508 abstract： Deep Reinforcement Learning (RL) is well known for being highly sensitive to hyperparameters, requiring practitioners substantial efforts to optimize them for the problem at hand. This also limits the applicability of RL in real-world scenarios. In recent years, the field of automated Reinforcement Learning (AutoRL) has grown in popularity by trying to address this issue. However, these approaches typically hinge on additional samples to select well-performing hyperparameters, hindering sample-efficiency and practicality. Furthermore, most AutoRL methods are heavily based on already existing AutoML methods, which were originally developed neglecting the additional challenges inherent to RL due to its non-stationarities. In this work, we propose a new approach for AutoRL, called *Adaptive Q-Network* (AdaQN), that is tailored to RL to take into account the non-stationarity of the optimization procedure without requiring additional samples. AdaQN learns several Q-functions, each one trained with different hyperparameters, which are updated online using the Q-function with the smallest approximation error as a shared target. Our selection scheme simultaneously handles different hyperparameters while coping with the non-stationarity induced by the RL optimization procedure and being orthogonal to any critic-based RL algorithm. We demonstrate that AdaQN is theoretically sound and empirically validate it in MuJoCo control problems and Atari 2600 games, showing benefits in sample-efficiency, overall performance, robustness to stochasticity and training stability. Our code is available at *https://github.com/theovincent/AdaDQN*.

684. Swing-by Dynamics in Concept Learning and Compositional Generalization

链接：https://iclr.cc/virtual/2025/poster/28151 abstract： Prior work has shown that text-conditioned diffusion models can learn to identify and manipulate primitive concepts underlying a compositional data-generating process, enabling generalization to entirely novel, out-of-distribution compositions. Beyond performance evaluations, these studies develop a rich empirical phenomenology of learning dynamics, showing that models generalize sequentially, respecting the compositional hierarchy of

the data-generating process. Moreover, concept-centric structures within the data significantly influence a model's speed of learning the ability to manipulate a concept.In this paper, we aim to better characterize these empirical results from a theoretical standpoint.Specifically, we propose an abstraction of prior work's compositional generalization problem by introducing a structured identity mapping (SIM) task, where a model is trained to learn the identity mapping on a Gaussian mixture with structurally organized centroids. We mathematically analyze the learning dynamics of neural networks trained on this SIM task and show that, despite its simplicity, SIM's learning dynamics capture and help explain key empirical observations on compositional generalization with diffusion models identified in prior work.Our theory also offers several new insights---e.g., we find a novel mechanism for non-monotonic learning dynamics of test loss in early phases of training.We validate our new predictions by training a text-conditioned diffusion model, bridging our simplified framework and complex generative models.Overall, this work establishes the SIM task as a meaningful theoretical abstraction of concept learning dynamics in modern generative models.

685. Single-agent Poisoning Attacks Suffice to Ruin Multi-Agent Learning

链接：https://iclr.cc/virtual/2025/poster/31035 abstract： We investigate the robustness of multi-agent learning in strongly monotone games with bandit feedback. While previous research has developed learning algorithms that achieve last-iterate convergence to the unique Nash equilibrium (NE) at a polynomial rate, we demonstrate that all such algorithms are vulnerable to adversaries capable of poisoning even a single agent's utility observations. Specifically, we propose an attacking strategy such that for any given time horizon T, the adversary can mislead any multi-agent learning algorithm to converge to a point other than the unique NE with a corruption budget that grows sublinearly in T. To further understand the inherent robustness of these algorithms, we characterize the fundamental trade-off between convergence speed and the maximum tolerable total utility corruptions for two example algorithms, including the state-of-the-art one. Our theoretical and empirical results reveal an intrinsic efficiency-robustness trade-off: the faster an algorithm converges, the more vulnerable it becomes to utility poisoning attacks. To the best of our knowledge, this is the first work to identify and characterize such a trade-off in the context of multi-agent learning.

686. Do Contemporary Causal Inference Models Capture Real-World Heterogeneity? Findings from a Large-Scale Benchmark

链接：https://iclr.cc/virtual/2025/poster/29711 abstract： We present unexpected findings from a large-scale benchmark study evaluating Conditional Average Treatment Effect (CATE) estimation algorithms. By running 16 modern CATE models across 43,200 datasets, we find that: (a) 62\% of CATE estimates have a higher Mean Squared Error (MSE) than a trivial zero-effect predictor, rendering them ineffective; (b) in datasets with at least one useful CATE estimate, 80\% still have higher MSE than a constant-effect model; and (c) Orthogonality-based models outperform other models only 30\% of the time, despite widespread optimism about their performance. These findings expose significant limitations in current CATE models and suggest ample opportunities for further research.Our findings stem from a novel application of \textit{observational sampling}, originally developed to evaluate Average Treatment Effect (ATE) estimates from observational methods with experiment data. To adapt observational sampling for CATE evaluation, we introduce a statistical parameter, Q, equal to MSE minus a constant and preserves the ranking of models by their MSE. We then derive a family of sample statistics, collectively called $\hat{Q}$, that can be computed from real-world data. We prove that $\hat{Q}$ is a consistent estimator of Q under mild technical conditions. When used in observational sampling, $\hat{Q}$ is unbiased and asymptotically selects the model with the smallest MSE. To ensure the benchmark reflects real-world heterogeneity, we handpick datasets where outcomes come from field rather than simulation. By combining the new observational sampling method, new statistics, and real-world datasets, the benchmark provides a unique perspective on CATE estimator performance and uncover gaps in capturing real-world heterogeneity.

687. A Differentiable Rank-Based Objective for Better Feature Learning

链接：https://iclr.cc/virtual/2025/poster/30039 abstract： In this paper, we leverage existing statistical methods to better understand feature learning from data. We tackle this by modifying the model-free variable selection method, Feature Ordering by Conditional Independence (FOCI), which is introduced in Azadkia & Chatterjee (2021). While FOCI is based on a non-parametric coefficient of conditional dependence, we introduce its parametric, differentiable approximation. With this approximate coefficient of correlation, we present a new algorithm called difFOCI, which is applicable to a wider range of machine learning problems thanks to its differentiable nature and learnable parameters. We present difFOCI in three contexts: (1) as a variable selection method with baseline comparisons to FOCI, (2) as a trainable model parametrized with a neural network, and (3) as a generic, widely applicable neural network regularizer, one that improves feature learning with better management of spurious correlations. We evaluate difFOCI on increasingly complex problems ranging from basic variable selection in toy examples to saliency map comparisons in convolutional networks. We then show how difFOCI can be incorporated in the context of fairness to facilitate classifications without relying on sensitive data.

688. Mixture of In-Context Prompters for Tabular PFNs

链接：https://iclr.cc/virtual/2025/poster/31124 abstract： Recent benchmarks find In-Context Learning (ICL) outperforms both deep learning and tree-based algorithms on small tabular datasets. However, on larger datasets, ICL for tabular learning suffers in both efficiency and effectiveness. In terms of efficiency, transformers incur linear space and quadratic time complexity w.r.t. context size. In terms of effectiveness, contexts at inference encounter distribution shift compared to contexts from pretraining. We propose MixturePFN, which extends Sparse Mixture of Experts to the state-of-the-art ICL for tabular learning model. Specifically, MixturePFN finetunes a specialized ICL expert on each cluster of tabular data and routes new test samples to

appropriate experts at inference. MixturePFN supports constant-size contexts by splitting large training datasets into more manageable clusters. MixturePFN addresses distribution shift by finetuning an expert on each training dataset cluster via bootstrapping. Extensive experimental results shows MixturePFN outperforms 19 baselines both in mean rank and as the Condorcet winner across 36 diverse tabular datasets under both accuracy and F1 score with statistical significance.

689. Learning Long Range Dependencies on Graphs via Random Walks

链接：https://iclr.cc/virtual/2025/poster/28595 abstract： Message-passing graph neural networks (GNNs) excel at capturing local relationships but struggle with long-range dependencies in graphs. In contrast, graph transformers (GTs) enable global information exchange but often oversimplify the graph structure by representing graphs as sets of fixed-length vectors. This work introduces a novel architecture that overcomes the shortcomings of both approaches by combining the long-range information of random walks with local message passing. By treating random walks as sequences, our architecture leverages recent advances in sequence models to effectively capture long-range dependencies within these walks. Based on this concept, we propose a framework that offers (1) more expressive graph representations through random walk sequences, (2) the ability to utilize any sequence model for capturing long-range dependencies, and (3) the flexibility by integrating various GNN and GT architectures. Our experimental evaluations demonstrate that our approach achieves competitive performance on 19 graph and node benchmark datasets, notably outperforming existing methods by up to 13\% on the PascalVoc-SP and COCO-SP datasets.Code: https://github.com/BorgwardtLab/NeuralWalker

690. Point-based Instance Completion with Scene Constraints

链接：https://iclr.cc/virtual/2025/poster/28501 abstract： Recent point-based object completion methods have demonstrated the ability to accurately recover the missing geometry of partially observed objects. However, these approaches are not well-suited for completing objects within a scene, as they do not consider known scene constraints (e.g., other observed surfaces) in their completions and further expect the partial input to be in a canonical coordinate system, which does not hold for objects within scenes. While instance scene completion methods have been proposed for completing objects within a scene, they lag behind point-based object completion methods in terms of object completion quality and still do not consider known scene constraints during completion. To overcome these limitations, we propose a point cloud-based instance completion model that can robustly complete objects at arbitrary scales and pose in the scene. To enable reasoning at the scene level, we introduce a sparse set of scene constraints represented as point clouds and integrate them into our completion model via a cross-attention mechanism. To evaluate the instance scene completion task on indoor scenes, we further build a new dataset called ScanWCF, which contains labeled partial scans as well as aligned ground truth scene completions that are watertight and collision-free. Through several experiments, we demonstrate that our method achieves improved fidelity to partial scans, higher completion quality, and greater plausibility over existing state-of-the-art methods.

691. Geometry-Aware Approaches for Balancing Performance and Theoretical Guarantees in Linear Bandits

链接：https://iclr.cc/virtual/2025/poster/29812 abstract： This paper is motivated by recent research in the d-dimensional stochastic linear bandit literature, which has revealed an unsettling discrepancy: algorithms like Thompson sampling and Greedy demonstrate promising empirical performance, yet this contrasts with their pessimistic theoretical regret bounds. The challenge arises from the fact that while these algorithms may perform poorly in certain problem instances, they generally excel in typical instances. To address this, we propose a new data-driven technique that tracks the geometric properties of the uncertainty ellipsoid around the main problem parameter. This methodology enables us to formulate a data-driven frequentist regret bound, which incorporates the geometric information, for a broad class of base algorithms, including Greedy, OFUL, and Thompson sampling. This result allows us to identify and ``course-correct'' problem instances in which the base algorithms perform poorly. The course-corrected algorithms achieve the minimax optimal regret of order $\tilde{\mathcal{O}}(d\sqrt{T})$ for a T-period decision-making scenario, effectively maintaining the desirable attributes of the base algorithms, including their empirical efficacy. We present simulation results to validate our findings using synthetic and real data.

692. Graph Neural Networks Can (Often) Count Substructures

链接：https://iclr.cc/virtual/2025/poster/28119 abstract： Message passing graph neural networks (GNNs) are known to have limited expressive power in their ability to distinguish some non-isomorphic graphs.Because of this, it is well known that they are unable to detect or count arbitrary graph substructures (i.e., solving the subgraph isomorphism problem), a task that is of great importance for several types of graph-structured data. However, we observe that GNNs are in fact able to count graph patterns quite accurately across several real-world graph datasets.Motivated by this observation, we provide an analysis of the subgraph-counting capabilities of GNNs beyond the worst case, deriving several sufficient conditions for GNNs to be able to count subgraphs and, more importantly, to be able to sample-efficiently learn to count subgraphs. Moreover, we develop novel dynamic programming algorithms for solving the subgraph isomorphism problem on restricted classes of pattern and target graphs, and show that message-passing GNNs can efficiently simulate these dynamic programs. Finally, we empirically validate that our sufficient conditions for GNNs to count subgraphs hold on many real-world datasets, providing a theoretically-grounded explanation to our motivating observations.

693. ICLR: In-Context Learning of Representations

链接：https://iclr.cc/virtual/2025/poster/28292 abstract： Recent work demonstrates that structured patterns in pretraining data influence how representations of different concepts are organized in a large language model's (LLM) internals, with such representations then driving downstream abilities. Given the open-ended nature of LLMs, e.g., their ability to in-context learn novel tasks, we ask whether models can flexibly alter their semantically grounded organization of concepts. Specifically, if we provide in-context exemplars wherein a concept plays a different role than what the pretraining data suggests, can models infer these novel semantics and reorganize representations in accordance with them? To answer this question, we define a toy "graph tracing" task wherein the nodes of the graph are referenced via concepts seen during training (e.g., apple, bird, etc.), and the connectivity of the graph is defined via some predefined structure (e.g., a square grid). Given exemplars that indicate traces of random walks on the graph, we analyze intermediate representations of the model and find that as the amount of context is scaled, there is a sudden re-organization of representations according to the graph's structure. Further, we find that when reference concepts have correlations in their semantics (e.g., Monday, Tuesday, etc.), the context-specified graph structure is still present in the representations, but is unable to dominate the pretrained structure. To explain these results, we analogize our task to energy minimization for a predefined graph topology, which shows getting non-trivial performance on the task requires for the model to infer a connected component. Overall, our findings indicate context-size may be an underappreciated scaling axis that can flexibly re-organize model representations, unlocking novel capabilities.

694. Optimized Multi-Token Joint Decoding With Auxiliary Model for LLM Inference

链接：https://iclr.cc/virtual/2025/poster/29226 abstract： Large language models (LLMs) have achieved remarkable success across diverse tasks, yet their inference processes are hindered by substantial time and energy demands due to single-token generation at each decoding step. While previous methods such as speculative decoding mitigate these inefficiencies by producing multiple tokens per step, each token is still generated by its single-token distribution,thereby enhancing speed without improving effectiveness. In contrast, our work simultaneously enhances inference speed and improves the output effectiveness. We consider multi-token joint decoding (MTJD), which generates multiple tokens from their joint distribution at each iteration, theoretically reducing perplexity and enhancing task performance. However, MTJD suffers from the high cost of sampling from the joint distribution of multiple tokens. Inspired by speculative decoding, we introduce multi-token assisted decoding (MTAD), a novel framework designed to accelerate MTJD. MTAD leverages a smaller auxiliary model to approximate the joint distribution of a larger model, incorporating a verification mechanism that not only ensures the accuracy of this approximation, but also improves thedecoding efficiency over conventional speculative decoding. Theoretically, we demonstrate that MTAD closely approximates exact MTJD with bounded error. Empirical evaluations using Llama-2 and OPT models ranging from 13B to 70B parameters across various tasks reveal that MTAD reduces perplexity by 21.2% and improves downstream performance compared to standard single-token sampling.Furthermore, MTAD achieves a 1.42× speed-up and consumes 1.54× less energy than conventional speculative decoding methods. These results highlight MTAD's ability to make multi-token joint decoding both effective and efficient, promoting more sustainable and high-performance deployment of LLMs.

695. MAGNet: Motif-Agnostic Generation of Molecules from Scaffolds

链接：https://iclr.cc/virtual/2025/poster/30958 abstract： Recent advances in machine learning for molecules exhibit great potential for facilitating drug discovery from in silico predictions.Most models for molecule generation rely on the decomposition of molecules into frequently occurring substructures (motifs), from which they generate novel compounds. While motif representations greatly aid in learning molecular distributions, such methods fail to represent substructures beyond their known motif set, posing a fundamental limitation for discovering novel compounds.To address this limitation and enhance structural expressivity, we propose to separate structure from features by abstracting motifs to scaffolds and, subsequently, allocating atom and bond types. To this end, we introduce a novel factorisation of the molecules' data distribution that considers the entire molecular context and facilitates learning adequate assignments of atoms and bonds to scaffolds. Complementary to this, we propose MAGNet, the first model to freely learn motifs. Importantly, we demonstrate that MAGNet's improved expressivity leads to molecules with more structural diversity and, at the same time, diverse atom and bond assignments.

696. Spectral Compressive Imaging via Unmixing-driven Subspace Diffusion Refinement

链接：https://iclr.cc/virtual/2025/poster/29715 abstract： Spectral Compressive Imaging (SCI) reconstruction is inherently ill-posed, offering multiple plausible solutions from a single observation. Traditional deterministic methods typically struggle to effectively recover high-frequency details. Although diffusion models offer promising solutions to this challenge, their application is constrained by the limited training data and high computational demands associated with multispectral images (MSIs), complicating direct training. To address these issues, we propose a novel Predict-and-unmixing-driven-Subspace-Refine framework (PSR-SCI). This framework begins with a cost-effective predictor that produces an initial, rough estimate of the MSI. Subsequently, we introduce a unmixing-driven reversible spectral embedding module that decomposes the MSI into subspace images and spectral coefficients. This decomposition facilitates the adaptation of pre-trained RGB diffusion models and focuses refinement processes on high-frequency details, thereby enabling efficient diffusion generation with minimal MSI data. Additionally, we design a high-dimensional guidance mechanism with imaging consistency to enhance the model's efficacy. The refined subspace image is then reconstructed back into an MSI using the reversible embedding, yielding the final MSI with full spectral resolution. Experimental results on the standard KAIST and zero-shot datasets NTIRE, ICVL, and Harvard show that PSR-SCI enhances visual quality and delivers PSNR and SSIM metrics comparable to existing diffusion, transformer, and deep unfolding techniques. This framework provides a robust alternative to traditional deterministic SCI reconstruction methods. Code

and models are available at https://github.com/SMARK2022/PSR-SCI.

697. How to Probe: Simple Yet Effective Techniques for Improving Post-hoc Explanations

链接：https://iclr.cc/virtual/2025/poster/30968 abstract： Post-hoc importance attribution methods are a popular tool for “explaining” Deep Neural Networks (DNNs) and are inherently based on the assumption that the explanations can be applied independently of how the models were trained. Contrarily, in this work we bring forward empirical evidence that challenges this very notion. Surprisingly, we discover a strong dependency on and demonstrate that the training details of a pre-trained model’s classification layer (<10% of model parameters) play a crucial role, much more than the pre-training scheme itself. This is of high practical relevance: (1) as techniques for pre-training models are becoming increasingly diverse, understanding the interplay between these techniques and attribution methods is critical; (2) it sheds light on an important yet overlooked assumption of post-hoc attribution methods which can drastically impact model explanations and how they are interpreted eventually. With this finding we also present simple yet effective adjustments to the classification layers, that can significantly enhance the quality of model explanations. We validate our findings across several visual pre-training frameworks (fully-supervised, self-supervised, contrastive vision-language training) and analyse how they impact explanations for a wide range of attribution methods on a diverse set of evaluation metrics.

698. As large as it gets – Studying Infinitely Large Convolutions via Neural Implicit Frequency Filters

链接：https://iclr.cc/virtual/2025/poster/31487 abstract： Recent work in neural networks for image classification has seen a strong tendency towards increasing the spatial context during encoding. Whether achieved through large convolution kernels or self-attention, models scale poorly with the increased spatial context, such that the improved model accuracy often comes at significant costs. In this paper, we propose a module for studying the effective filter size of convolutional neural networks (CNNs). To facilitate such a study, several challenges need to be addressed: (i) we need an effective means to train models with large filters (potentially as large as the input data) without increasing the number of learnable parameters, (ii) the employed convolution operation should be a plug-and-play module that can replace conventional convolutions in a CNN and allow for an efficient implementation in current frameworks, (iii) the study of filter sizes has to be decoupled from other aspects such as the network width or the number of learnable parameters, and (iv) the cost of the convolution operation itself has to remain manageable i.e.~we can not naïvely increase the size of the convolution kernel. To address these challenges, we propose to learn the frequency representations of filter weights as neural implicit functions, such that the better scalability of the convolution in the frequency domain can be leveraged. Additionally, due to the implementation of the proposed neural implicit function, even large and expressive spatial filters can be parameterized by only a few learnable weights. Interestingly, our analysis shows that, although the proposed networks could learn very large convolution kernels, the learned filters are well localized and relatively small in practice when transformed from the frequency to the spatial domain. We anticipate that our analysis of individually optimized filter sizes will allow for more efficient, yet effective, models in the future. Our code is available at https://github.com/GeJulia/NIFF .

699. Tracking the Copyright of Large Vision-Language Models through Parameter Learning Adversarial Images

链接：https://iclr.cc/virtual/2025/poster/30074 abstract： Large vision-language models (LVLMs) have demonstrated remarkable image understanding and dialogue capabilities, allowing them to handle a variety of visual question answering tasks. However, their widespread availability raises concerns about unauthorized usage and copyright infringement, where users or individuals can develop their own LVLMs by fine-tuning published models. In this paper, we propose a novel method called Parameter Learning Attack (PLA) for tracking the copyright of LVLMs without modifying the original model. Specifically, we construct adversarial images through targeted attacks against the original model, enabling it to generate specific outputs. To ensure these attacks remain effective on potential fine-tuned models to trigger copyright tracking, we allow the original model to learn the trigger images by updating parameters in the opposite direction during the adversarial attack process. Notably, the proposed method can be applied after the release of the original model, thus not affecting the model’s performance and behavior. To simulate real-world applications, we fine-tune the original model using various strategies across diverse datasets, creating a range of models for copyright verification. Extensive experiments demonstrate that our method can more effectively identify the original copyright of fine-tuned models compared to baseline methods. Therefore, this work provides a powerful tool for tracking copyrights and detecting unlicensed usage of LVLMs.

700. Towards Scalable Exact Machine Unlearning Using Parameter-Efficient Fine-Tuning

链接：https://iclr.cc/virtual/2025/poster/28344 abstract： Machine unlearning is the process of efficiently removing the influence of a training data instance from a trained machine learning model without retraining it from scratch. A popular subclass of unlearning approaches is exact machine unlearning, which focuses on techniques that explicitly guarantee the removal of the influence of a data instance from a model. Exact unlearning approaches use a machine learning model in which individual components are trained on disjoint subsets of the data. During deletion, exact unlearning approaches only retrain the affected

components rather than the entire model. While existing approaches reduce retraining costs, it can still be expensive for an organization to retrain a model component as it requires halting a system in production, which leads to service failure and adversely impacts customers. To address these challenges, we introduce an exact unlearning framework -- Sequence-aware Sharded Sliced Training (S3T), which is designed to enhance the deletion capabilities of an exact unlearning system while minimizing the impact on model's performance. At the core of S3T, we utilize a lightweight parameter-efficient fine-tuning approach that enables parameter isolation by sequentially training layers with disjoint data slices. This enables efficient unlearning by simply deactivating the layers affected by data deletion. Furthermore, to reduce the retraining cost and improve model performance, we train the model on multiple data sequences, which allows S3T to handle an increased number of deletion requests. Both theoretically and empirically, we demonstrate that S3T attains superior deletion capabilities and enhanced performance compared to baselines across a wide range of settings.

701. Near, far: Patch-ordering enhances vision foundation models' scene understanding

链接：https://iclr.cc/virtual/2025/poster/29671 abstract： We introduce NeCo: Patch Neighbor Consistency, a novel self-supervised training loss that enforces patch-level nearest neighbor consistency across a student and teacher model. Compared to contrastive approaches that only yield binary learning signals, i.e. "attract" and "repel", this approach benefits from the more fine-grained learning signal of sorting spatially dense features relative to reference patches. Our method leverages differentiable sorting applied on top of pretrained representations, such as DINOv2-registers to bootstrap the learning signal and further improve upon them. This dense post-pretraining leads to superior performance across various models and datasets, despite requiring only 19 hours on a single GPU. This method generates high-quality dense feature encoders and establishes several new state-of-the-art results such as +2.3 % and +4.2% for non-parametric in-context semantic segmentation on ADE20k and Pascal VOC, +1.6% and +4.8% for linear segmentation evaluations on COCO-Things and -Stuff and improvements in the 3D understanding of multi-view consistency on SPair-71k, by more than 1.5%.

702. Unifying Causal Representation Learning with the Invariance Principle

链接：https://iclr.cc/virtual/2025/poster/28503 abstract： Causal representation learning (CRL) aims at recovering latent causal variables from high-dimensional observations to solve causal downstream tasks, such as predicting the effect of new interventions or more robust classification. A plethora of methods have been developed, each tackling carefully crafted problem settings that lead to different types of identifiability. These different settings are widely assumed to be important because they are often linked to different rungs of Pearl's causal hierarchy, even though this correspondence is not always exact. This work shows that instead of strictly conforming to this hierarchical mapping, many causal representation learning approaches methodologically align their representations with inherent data symmetries. Identification of causal variables is guided by invariance principles that are not necessarily causal. This result allows us to unify many existing approaches in a single method that can mix and match different assumptions, including non-causal ones, based on the invariance relevant to the problem at hand. It also significantly benefits applicability, which we demonstrate by improving treatment effect estimation on real-world high-dimensional ecological data. Overall, this paper clarifies the role of causal assumptions in the discovery of causal variables and shifts the focus to preserving data symmetries.

703. Scalable Mechanistic Neural Networks

链接：https://iclr.cc/virtual/2025/poster/29818 abstract： We propose Scalable Mechanistic Neural Network (S-MNN), an enhanced neural network framework designed for scientific machine learning applications involving long temporal sequences. By reformulating the original Mechanistic Neural Network (MNN) (Pervez et al., 2024), we reduce the computational time and space complexities from cubic and quadratic with respect to the sequence length, respectively, to linear. This significant improvement enables efficient modeling of long-term dynamics without sacrificing accuracy or interpretability. Extensive experiments demonstrate that S-MNN matches the original MNN in precision while substantially reducing computational resources. Consequently, S-MNN can drop-in replace the original MNN in applications, providing a practical and efficient tool for integrating mechanistic bottlenecks into neural network models of complex dynamical systems. Source code is available at https://github.com/IST-DASLab/ScalableMNN.

704. Deep Signature: Characterization of Large-Scale Molecular Dynamics

链接：https://iclr.cc/virtual/2025/poster/27774 abstract： Understanding protein dynamics are essential for deciphering protein functional mechanisms and developing molecular therapies. However, the complex high-dimensional dynamics and interatomic interactions of biological processes pose significant challenge for existing computational techniques. In this paper, we approach this problem for the first time by introducing Deep Signature, a novel computationally tractable framework that characterizes complex dynamics and interatomic interactions based on their evolving trajectories. Specifically, our approach incorporates soft spectral clustering that locally aggregates cooperative dynamics to reduce the size of the system, as well as signature transform that collects iterated integrals to provide a global characterization of the non-smooth interactive dynamics. Theoretical analysis demonstrates that Deep Signature exhibits several desirable properties, including invariance to translation, near invariance to rotation, equivariance to permutation of atomic coordinates, and invariance under time reparameterization. Furthermore, experimental results on three benchmarks of biological processes verify that our approach can achieve superior performance compared to baseline methods.

705. Distributional Associations vs In-Context Reasoning: A Study of Feed-forward and Attention Layers

链接：https://iclr.cc/virtual/2025/poster/29376 abstract： Large language models have been successful at tasks involving basic forms of in-context reasoning, such as generating coherent language, as well as storing vast amounts of knowledge. At the core of the Transformer architecture behind such models are feed-forward and attention layers, which are often associated to knowledge and reasoning, respectively. In this paper, we study this distinction empirically and theoretically in a controlled synthetic setting where certain next-token predictions involve both distributional and in-context information. We find that feed-forward layers tend to learn simple distributional associations such as bigrams, while attention layers focus on in-context reasoning. Our theoretical analysis identifies the noise in the gradients as a key factor behind this discrepancy. Finally, we illustrate how similar disparities emerge in pre-trained models through ablations on the Pythia model family on simple reasoning tasks.

706. Quality over Quantity in Attention Layers: When Adding More Heads Hurts

链接：https://iclr.cc/virtual/2025/poster/27747 abstract： Attention-based mechanisms are widely used in machine learning, most prominently in transformers. However, hyperparameters such as the number of attention heads and the attention rank (i.e., the query/key dimension) are set nearly the same way in all realizations of this architecture, without theoretical justification. In this paper, we prove that the rank can have a dramatic effect on the representational capacity of attention. This effect persists even when the number of heads and the parameter count are very large. Specifically, we present a simple and natural target function based on nearest neighbor search that can be represented using a single full-rank attention head for any sequence length, but that cannot be approximated by a low-rank attention layer even on short sequences unless the number of heads is exponential in the embedding dimension. Thus, for this target function, rank is what determines an attention layer's power. We show that, for short sequences, using multiple layers allows the target to be approximated by low-rank attention; for long sequences, we conjecture that full-rank attention is necessary regardless of depth. Finally, we present experiments with standard multilayer transformers that validate our theoretical findings. They demonstrate that, because of how standard transformer implementations set the rank, increasing the number of attention heads can severely decrease accuracy on certain tasks.

707. PICASO: Permutation-Invariant Context Composition with State Space Models

链接：https://iclr.cc/virtual/2025/poster/30782 abstract： Providing Large Language Models with relevant contextual knowledge at inference time has been shown to greatly improve the quality of their generations. This is often achieved by prepending informative passages of text, or 'contexts', retrieved from external knowledge bases to their input. However, processing additional contexts online incurs significant computation costs that scale with their length. State Space Models (SSMs) offer a promising solution by allowing a database of contexts to be mapped onto fixed-dimensional states from which to start the generation. A key challenge arises when attempting to leverage information present across multiple contexts, since there is no straightforward way to condition generation on multiple independent states in existing SSMs. To address this, we leverage a simple mathematical relation derived from SSM dynamics to compose multiple states into one that efficiently approximates the effect of concatenating raw context tokens. Since the temporal ordering of contexts can often be uninformative, we enforce permutation-invariance by efficiently averaging states obtained via our composition algorithm across all possible context orderings. We evaluate our resulting method on WikiText and MSMARCO in both zero-shot and fine-tuned settings, and show that we can match the strongest performing baseline while enjoying on average $5.4\times$ speedup.

708. InstaSHAP: Interpretable Additive Models Explain Shapley Values Instantly

链接：https://iclr.cc/virtual/2025/poster/28549 abstract： In recent years, the Shapley value and SHAP explanations have emerged as oneof the most dominant paradigms for providing post-hoc explanations of blackbox models. Despite their well-founded theoretical properties, many recent workshave focused on the limitations in both their computational efficiency and theirrepresentation power. The underlying connection with additive models, however,is left critically under-emphasized in the current literature. In this work, we findthat a variational perspective linking GAM models and SHAP explanations is ableto provide deep insights into nearly all recent developments. In light of this connection, we borrow in the other direction to develop a new method to train interpretable GAM models which are automatically purified to compute the Shapleyvalue in a single forward pass. Finally, we provide theoretical results showing thelimited representation power of GAM models is the same Achilles’ heel existingin SHAP and discuss the implications for SHAP’s modern usage in CV and NLP.

709. A Large-scale Dataset and Benchmark for Commuting Origin-Destination Flow Generation

链接：https://iclr.cc/virtual/2025/poster/29354 abstract： Commuting Origin-Destination~(OD) flows are critical inputs for urban planning and transportation, providing crucial information about the population residing in one region and working in another

within an interested area. Due to the high cost of data collection, researchers have developed physical and computational models to generate commuting OD flows using readily available urban attributes, such as sociodemographics and points of interest, for cities lacking historical OD flows \textemdash commuting OD flow generation. Existing works developed models based on different techniques and achieved improvement on different datasets with different evaluation metrics, which hinderes establishing a unified standard for comparing model performance. To bridge this gap, we introduce a large-scale dataset containing commuting OD flows for 3,333 areas including a wide range of urban environments around the United States. Based on that, we benchmark widely used models for commuting OD flow generation. We surprisingly find that the network-based generative models achieve the optimal performance in terms of both precision and generalization ability, which may inspire new research directions of graph generative modeling in this field. The dataset and benchmark are available at https://anonymous.4open.science/r/CommutingODGen-Dataset-0D4C/.

710. VEDIT: Latent Prediction Architecture For Procedural Video Representation Learning

链接：https://iclr.cc/virtual/2025/poster/30010 abstract： Procedural video representation learning is an active research area where the objective is to learn an agent which can anticipate and forecast the future given the present video input, typically in conjunction with textual annotations. Prior works often rely on large-scale pretraining of visual encoders and prediction models with language supervision. However, the necessity and effectiveness of extending compute intensive pretraining to learn video clip sequences with noisy text supervision have not yet been fully validated by previous works. In this work, we show that a strong off-the-shelf frozen pretrained visual encoder, along with a well designed prediction model, can achieve state-of-the-art (SoTA) performance in forecasting and procedural planning without the need for pretraining the prediction model, nor requiring additional supervision from language or ASR. Instead of learning representations from pixel space, our method utilizes the latent embedding space of publicly available vision encoders. By conditioning on frozen clip-level embeddings from observed steps to predict the actions of unseen steps, our prediction model is able to learn robust representations for forecasting through iterative denoising —leveraging the recent advances in diffusion transformers (Peebles & Xie, 2023). Empirical studies over a total of five procedural learning tasks across four datasets (NIV, CrossTask, COIN and Ego4D-v2) show that our model advances the strong baselines in long-horizon action anticipation (+2.6% in Verb ED@20, +3.1% in Noun ED@20), and significantly improves the SoTA in step forecasting (+5.0%), task classification (+3.8%), and procedure planning tasks (up to +2.28% in success rate, +3.39% in mAcc, and +0.90% in mIoU).

711. A Multiscale Frequency Domain Causal Framework for Enhanced Pathological Analysis

链接：https://iclr.cc/virtual/2025/poster/30857 abstract： Multiple Instance Learning (MIL) in digital pathology Whole Slide Image (WSI) analysis has shown significant progress. However, due to data bias and unobservable confounders, this paradigm still faces challenges in terms of performance and interpretability. Existing MIL methods might identify patches that do not have true diagnostic significance, leading to false correlations, and experience difficulties in integrating multi-scale features and handling unobservable confounders. To address these issues, we propose a new Multi-Scale Frequency Domain Causal framework (MFC). This framework employs an adaptive memory module to estimate the overall data distribution through multi-scale frequency-domain information during training and simulates causal interventions based on this distribution to mitigate confounders in pathological diagnosis tasks. The framework integrates the Multi-scale Spatial Representation Module (MSRM), Frequency Domain Structure Representation Module (FSRM), and Causal Memory Intervention Module (CMIM) to enhance the model's performance and interpretability. Furthermore, the plug-and-play nature of this framework allows it to be broadly applied across various models. Experimental results on Camelyon16 and TCGA-NSCLC dataset show that, compared to previous work, our method has significantly improved accuracy and generalization ability, providing a new theoretical perspective for medical image analysis and potentially advancing the field further. The code will be released at https://github.com/WissingChen/MFC-MIL.

712. Bidirectional Decoding: Improving Action Chunking via Guided Test-Time Sampling

链接：https://iclr.cc/virtual/2025/poster/28245 abstract： Predicting and executing a sequence of actions without intermediate replanning, known as action chunking, is increasingly used in robot learning from human demonstrations. Yet, its effects on the learned policy remain inconsistent: some studies find it crucial for achieving strong results, while others observe decreased performance. In this paper, we first dissect how action chunking impacts the divergence between a learner and a demonstrator. We find that action chunking allows the learner to better capture the temporal dependencies in demonstrations but at the cost of reduced reactivity to unexpected states. To address this tradeoff, we propose Bidirectional Decoding (BID), a test-time inference algorithm that bridges action chunking with closed-loop adaptation. At each timestep, BID samples multiple candidate predictions and searches for the optimal one based on two criteria: (i) backward coherence, which favors samples that align with previous decisions; (ii) forward contrast, which seeks samples of high likelihood for future plans. By coupling decisions within and across action chunks, BID promotes both long-term consistency and short-term reactivity. Experimental results show that our method boosts the performance of two state-of-the-art generative policies across seven simulation benchmarks and two real-world tasks. Code and videos are available at https://bid-robot.github.io.

713. Training-Free Activation Sparsity in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28994 abstract： Activation sparsity can enable practical inference speedups in large language models (LLMs) by reducing the compute and memory-movement required for matrix multiplications during the forward pass. However, existing methods face limitations that inhibit widespread adoption. Some approaches are tailored towards older models with ReLU-based sparsity, while others require extensive continued pre-training on up to hundreds of billions of tokens. This paper describes TEAL (Training-Free Activation Sparsity in LLMs), a simple training-free method that applies magnitude-based activation sparsity to hidden states throughout the entire model. TEAL achieves 40-50\% model-wide sparsity with minimal performance degradation across Llama-2, Llama-3, and Mistral families, with sizes varying from 7B to 70B. We improve existing sparse kernels and demonstrate wall-clock decoding speed-ups of up to 1.53× and 1.8× at 40\% and 50\% model-wide sparsity. TEAL is compatible with weight quantization, enabling further efficiency gains.

714. Trust or Escalate: LLM Judges with Provable Guarantees for Human Agreement

链接：https://iclr.cc/virtual/2025/poster/29482 abstract： We present a principled approach to provide LLM-based evaluation with a rigorous guarantee of human agreement. We first propose that a reliable evaluation method should not uncritically rely on model preferences for pairwise evaluation, but rather assess the confidence of judge models and selectively decide when to trust its judgement. We then show that under this selective evaluation framework, human agreement can be provably guaranteed---such that the model evaluation aligns with that of humans to a user-specified agreement level. As part of our framework, we also introduce Simulated Annotators, a novel confidence estimation method that significantly improves judge calibration and thus enables high coverage of evaluated instances. Finally, we propose Cascaded Selective Evaluation, where we use cheaper models as initial judges and escalate to stronger models only when necessary---again, while still providing a provable guarantee of human agreement. Experimental results show that Cascaded Selective Evaluation guarantees strong alignment with humans, far beyond what LLM judges could achieve without selective evaluation. For example, on a subset of Chatbot Arena where GPT-4 almost never achieves 80% human agreement, our method, even while employing substantially cost-effective models such as Mistral-7B, guarantees over 80% human agreement with almost 80% test coverage.

715. LLMs' Potential Influences on Our Democracy: Challenges and Opportunities

链接：https://iclr.cc/virtual/2025/poster/31347 abstract： With growing research and attention on LLMs' potential influence on political discourse and democratic processes, this blog post discusses the path forward and proposes future research questions in four broad areas: (1) evaluation of LLM political leanings, (2) understanding LLMs' influence on our democracy, (3) better policy frameworks for AI development, and (4) technical solutions to adjust or mitigate political leanings. As LLMs become increasingly integrated into society, continued investigation of how they will reshape democracy is essential to maximize their benefits while minimizing risks to democratic processes.

716. DailyDilemmas: Revealing Value Preferences of LLMs with Quandaries of Daily Life

链接：https://iclr.cc/virtual/2025/poster/29764 abstract： As users increasingly seek guidance from LLMs for decision-making in daily life, many of these decisions are not clear-cut and depend significantly on the personal values and ethical standards of people. We present DailyDilemmas, a dataset of 1,360 moral dilemmas encountered in everyday life. Each dilemma presents two possible actions, along with affected parties and relevant human values for each action. Based on these dilemmas, we gather a repository of human values covering diverse everyday topics, such as interpersonal relationships, workplace, and environmental issues. With DailyDilemmas, we evaluate LLMs on these dilemmas to determine what action they will choose and the values represented by these action choices. Then, we analyze values through the lens of five theoretical frameworks inspired by sociology, psychology, and philosophy, including the World Values Survey, Moral Foundations Theory, Maslow's Hierarchy of Needs, Aristotle's Virtues, and Plutchik's Wheel of Emotions. For instance, we find LLMs are most aligned with self-expression over survival in World Values Survey and care over loyalty in Moral Foundations Theory. Interestingly, we find substantial preference differences in models for some core values. For example, for truthfulness, Mixtral-8x7B neglects it by 9.7% while GPT-4-turbo selects it by 9.4%. We also study the recent guidance released by OpenAI (ModelSpec), and Anthropic (Constitutional AI) to understand how their designated principles reflect their models' actual value prioritization when facing nuanced moral reasoning in daily-life settings. Finally, we find that end users cannot effectively steer such prioritization using system prompts.

717. Magpie: Alignment Data Synthesis from Scratch by Prompting Aligned LLMs with Nothing

链接：https://iclr.cc/virtual/2025/poster/29730 abstract： High-quality instruction data is critical for aligning large language models (LLMs). Although some models, such as Llama-3-Instruct, have open weights, their alignment data remain private, which hinders the democratization of AI. High human labor costs and a limited, predefined scope for prompting prevent existing open-

source data creation methods from scaling effectively, potentially limiting the diversity and quality of public alignment datasets. Is it possible to synthesize high-quality instruction data at scale by extracting it directly from an aligned LLM? We present a self-synthesis method for generating large-scale alignment data named Magpie. Our key observation is that aligned LLMs like Llama-3-Instruct can generate a user query when we input only the pre-query templates up to the position reserved for user messages, thanks to their auto-regressive nature. We use this method to prompt Llama-3-Instruct and generate 4 million instructions along with their corresponding responses. We further introduce extensions of Magpie for filtering, generating multi-turn, preference optimization, domain-specific and multilingual datasets. We perform a comprehensive analysis of the Magpie-generated data. To compare Magpie-generated data with other public instruction datasets (e.g., ShareGPT, WildChat, Evol-Instruct, UltraChat, OpenHermes, Tulu-V2-Mix, GenQA), we fine-tune Llama-3-8B-Base with each dataset and evaluate the performance of the fine-tuned models. Our results indicate that using Magpie for supervised fine-tuning (SFT) solely can surpass the performance of previous public datasets utilized for both SFT and preference optimization, such as direct preference optimization with UltraFeedback. We also show that in some tasks, models supervised fine-tuned with Magpie perform comparably to the official Llama-3-8B-Instruct, despite the latter being enhanced with 10 million data points through SFT and subsequent preference optimization. This advantage is evident on alignment benchmarks such as AlpacaEval, ArenaHard, and WildBench.

718. Active Learning for Neural PDE Solvers

链接：https://iclr.cc/virtual/2025/poster/27803 abstract： Solving partial differential equations (PDEs) is a fundamental problem in engineering and science. While neural PDE solvers can be more efficient than established numerical solvers, they often require large amounts of training data that is costly to obtain. Active learning (AL) could help surrogate models reach the same accuracy with smaller training sets by querying classical solvers with more informative initial conditions and PDE parameters. While AL is more common in other domains, it has yet to be studied extensively for neural PDE solvers. To bridge this gap, we introduce AL4PDE, a modular and extensible active learning benchmark. It provides multiple parametric PDEs and state-of-the-art surrogate models for the solver-in-the-loop setting, enabling the evaluation of existing and the development of new AL methods for PDE solving. We use the benchmark to evaluate batch active learning algorithms such as uncertainty- and feature-based methods. We show that AL reduces the average error by up to 71\% compared to random sampling and significantly reduces worst-case errors. Moreover, AL generates similar datasets across repeated runs, with consistent distributions over the PDE parameters and initial conditions. The acquired datasets are reusable, providing benefits for surrogate models not involved in the data generation.

719. Utilitarian Algorithm Configuration for Infinite Parameter Spaces

链接：https://iclr.cc/virtual/2025/poster/31520 abstract： Utilitarian algorithm configuration is a general-purpose technique for automatically searching the parameter space of a given algorithm to optimize its performance, as measured by a given utility function, on a given set of inputs. Recently introduced utilitarian configuration procedures offer optimality guarantees about the returned parameterization while provably adapting to the hardness of the underlying problem. However, the applicability of these approaches is severely limited by the fact that they only search a finite, relatively small set of parameters. They cannot effectively search the configuration space of algorithms with continuous or uncountable parameters. In this paper we introduce a new procedure, which we dub COUP (Continuous, Optimistic Utilitarian Procrastination). COUP is designed to search infinite parameter spaces efficiently to find good configurations quickly. Furthermore, COUP maintains the theoretical benefits of previous utilitarian configuration procedures when applied to finite parameter spaces but is significantly faster, both provably and experimentally.

720. WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild

链接：https://iclr.cc/virtual/2025/poster/29940 abstract： We introduce WildBench, an automated evaluation framework designed to benchmark large language models (LLMs) using challenging, real-world user queries. WildBench consists of 1,024 tasks carefully selected from over one million human-chatbot conversation logs. For automated evaluation with WildBench, we have developed two metrics, WB-Reward and WB-Score, which are computable using advanced LLMs such as GPT-4-turbo. WildBench evaluation uses task-specific checklists to evaluate model outputs systematically and provides structured explanations that justify the scores and comparisons, resulting in more reliable and interpretable automatic judgments. WB-Reward employs fine-grained pairwise comparisons between model responses, generating five potential outcomes: much better, slightly better, slightly worse, much worse, or a tie. Unlike previous evaluations that employed a single baseline model, we selected three baseline models at varying performance levels to ensure a comprehensive pairwise evaluation. Additionally, we propose a simple method to mitigate length bias, by converting outcomes of “slightly better/worse” to “tie” if the winner response exceeds the loser one by more than K characters. WB-Score evaluates the quality of model outputs individually, making it a fast and cost-efficient evaluation metric. WildBench results demonstrate a strong correlation with the human-voted Elo ratings from Chatbot Arena on hard tasks. Specifically, WB-Reward achieves a Pearson correlation of 0.98 with top-ranking models. Additionally, WB-Score reaches 0.95, surpassing both ArenaHard’s 0.91 and AlpacaEval2.0’s 0.89 for length-controlled win rates, as well as the 0.87 for regular win rates.

721. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models

链接：https://iclr.cc/virtual/2025/poster/31514 abstract： Language models demonstrate both quantitative improvement and new qualitative capabilities with increasing scale. Despite their potentially transformative impact, these new capabilities are as yet poorly characterized. In order to inform future research, prepare for disruptive new model capabilities, and ameliorate socially harmful effects, it is vital that we understand the present and near-future capabilities and limitations of language models. To address this challenge, we introduce the Beyond the Imitation Game benchmark (BIG- bench). BIG-bench currently consists of 204 tasks, contributed by 450 authors across 132 institutions. Task topics are diverse, drawing problems from linguistics, childhood develop- ment, math, common-sense reasoning, biology, physics, social bias, software development, and beyond. BIG-bench focuses on tasks that are believed to be beyond the capabilities of current language models. We evaluate the behavior of OpenAI's GPT models, Google- internal dense transformer architectures, and Switch-style sparse transformers on BIG-bench, across model sizes spanning millions to hundreds of billions of parameters. In addition, a team of human expert raters performed all tasks in order to provide a strong baseline. Findings include: model performance and calibration both improve with scale, but are poor in absolute terms (and when compared with rater performance); performance is remarkably similar across model classes, though with benefits from sparsity; tasks that improve gradually and predictably commonly involve a large knowledge or memorization component, whereas tasks that exhibit "breakthrough" behavior at a critical scale often involve multiple steps or components, or brittle metrics; social bias typically increases with scale in settings with ambiguous context, but this can be improved with prompting.

722. X-Fi: A Modality-Invariant Foundation Model for Multimodal Human Sensing

链接：https://iclr.cc/virtual/2025/poster/29125 abstract： Human sensing, which employs various sensors and advanced deep learning technologies to accurately capture and interpret human body information, has significantly impacted fields like public security and robotics. However, current human sensing primarily depends on modalities such as cameras and LiDAR, each of which has its own strengths and limitations. Furthermore, existing multimodal fusion solutions are typically designed for fixed modality combinations, requiring extensive retraining when modalities are added or removed for diverse scenarios. In this paper, we propose a modality-invariant foundation model for all modalities, X-Fi, to address these issues. X-Fi enables the independent or combinatory use of sensor modalities without additional training by utilizing a transformer structure to accommodate variable input sizes and incorporating a novel "X-fusion" mechanism to preserve modality-specific features during multimodal integration. This approach not only enhances adaptability but also facilitates the learning of complementary features across modalities. Extensive experiments conducted on the MM-Fi and XRF55 datasets, employing six distinct modalities, demonstrate that X-Fi achieves state-of-the-art performance in human pose estimation (HPE) and human activity recognition (HAR) tasks. The findings indicate that our proposed model can efficiently support a wide range of human sensing applications, ultimately contributing to the evolution of scalable, multimodal sensing technologies.

723. Durable Quantization Conditioned Misalignment Attack on Large Language Models

链接：https://iclr.cc/virtual/2025/poster/31042 abstract： As large language models (LLMs) are increasingly deployed on resource-constrained edge devices, quantization techniques have been widely adopted to reduce model size and computational requirements. However, this process can expose models to new vulnerabilities. In this work, we introduce the Quantization Conditioned Misalignment (Q-Misalign) attack, a novel threat in which safety misalignment remains dormant in a full-precision LLM but becomes exploitable post-quantization. We demonstrate that our Q-Misalign attack effectively bypasses safety mechanisms and enables the generation of harmful content in quantized models while maintaining full-precision performance. Furthermore, we propose a contrastive task vector-based approach to enhance attack durability, ensuring that vulnerabilities persist even after downstream fine-tuning. Experimental results show that Q-Misalign attack significantly increases jailbreak success rates in quantized models, while preserving model utility and safety alignment in full precision. Our findings highlight a critical gap in current LLM safety measures and call for more robust defenses in quantization-aware scenarios.

724. Beyond Sequence: Impact of Geometric Context for RNA Property Prediction

链接：https://iclr.cc/virtual/2025/poster/30672 abstract： Accurate prediction of RNA properties, such as stability and interactions, is crucial for advancing our understanding of biological processes and developing RNA-based therapeutics. RNA structures can be represented as 1D sequences, 2D topological graphs, or 3D all-atom models, each offering different insights into its function. Existing works predominantly focus on 1D sequence-based models, which overlook the geometric context provided by 2D and 3D geometries. This study presents the first systematic evaluation of incorporating explicit 2D and 3D geometric information into RNA property prediction, considering not only performance but also real-world challenges such as limited data availability, partial labeling, sequencing noise, and computational efficiency. To this end, we introduce a newly curated set of RNA datasets with enhanced 2D and 3D structural annotations, providing a resource for model evaluation on RNA data. Our findings reveal that models with explicit geometry encoding generally outperform sequence-based models, with an average prediction RMSE reduction of around 12% across all various RNA tasks and excelling in low-data and partial labeling regimes, underscoring the value of explicitly incorporating geometric context. On the other hand, geometry-unaware sequence-based models are more robust under sequencing noise but often require around 2-5x training data to match the performance of geometry-aware models. Our study offers further insights into the trade-offs between different RNA representations in practical applications and addresses a significant gap in evaluating deep learning models for RNA tasks.

725. LLaRA: Supercharging Robot Learning Data for Vision-Language Policy

链接：https://iclr.cc/virtual/2025/poster/28695 abstract： Vision Language Models (VLMs) have recently been leveraged to generate robotic actions, forming Vision-Language-Action (VLA) models. However, directly adapting a pretrained VLM for robotic control remains challenging, particularly when constrained by a limited number of robot demonstrations. In this work, we introduce LLaRA: Large Language and Robotics Assistant, a framework that formulates robot action policy as visuo-textual conversations and enables an efficient transfer of a pretrained VLM into a powerful VLA, motivated by the success of visual instruction tuning in Computer Vision. First, we present an automated pipeline to generate conversation-style instruction tuning data for robots from existing behavior cloning datasets, aligning robotic actions with image pixel coordinates. Further, we enhance this dataset in a self-supervised manner by defining six auxiliary tasks, without requiring any additional action annotations. We show that a VLM finetuned with a limited amount of such datasets can produce meaningful action decisions for robotic control. Through experiments across multiple simulated and real-world tasks, we demonstrate that LLaRA achieves state-of-the-art performance while preserving the generalization capabilities of large language models. The code, datasets, and pretrained models are available at https://github.com/LostXine/LLaRA.

726. Epistemic Monte Carlo Tree Search

链接：https://iclr.cc/virtual/2025/poster/29531 abstract： The AlphaZero/MuZero (A/MZ) family of algorithms has achieved remarkable success across various challenging domains by integrating Monte Carlo Tree Search (MCTS) with learned models. Learned models introduce epistemic uncertainty, which is caused by learning from limited data and is useful for exploration in sparse reward environments. MCTS does not account for the propagation of this uncertainty however. To address this, we introduce Epistemic MCTS (EMCTS): a theoretically motivated approach to account for the epistemic uncertainty in search and harness the search for deep exploration. In the challenging sparse-reward task of writing code in the Assembly language SUBLEQ, AZ paired with our method achieves significantly higher sample efficiency over baseline AZ. Search with EMCTS solves variations of the commonly used hard-exploration benchmark Deep Sea - which baseline A/MZ are practically unable to solve - much faster than an otherwise equivalent method that does not use search for uncertainty estimation, demonstrating significant benefits from search for epistemic uncertainty estimation.

727. Advantage Alignment Algorithms

链接：https://iclr.cc/virtual/2025/poster/29701 abstract： Artificially intelligent agents are increasingly being integrated into human decision-making: from large language model (LLM) assistants to autonomous vehicles. These systems often optimize their individual objective, leading to conflicts, particularly in general-sum games where naive reinforcement learning agents empirically converge to Pareto-suboptimal Nash equilibria. To address this issue, opponent shaping has emerged as a paradigm for finding socially beneficial equilibria in general-sum games. In this work, we introduce Advantage Alignment, a family of algorithms derived from first principles that perform opponent shaping efficiently and intuitively. We achieve this by aligning the advantages of interacting agents, increasing the probability of mutually beneficial actions when their interaction has been positive. We prove that existing opponent shaping methods implicitly perform Advantage Alignment. Compared to these methods, Advantage Alignment simplifies the mathematical formulation of opponent shaping, reduces the computational burden and extends to continuous action domains. We demonstrate the effectiveness of our algorithms across a range of social dilemmas, achieving state-of-the-art cooperation and robustness against exploitation.

728. Solving hidden monotone variational inequalities with surrogate losses

链接：https://iclr.cc/virtual/2025/poster/31004 abstract： Deep learning has proven to be effective in a wide variety of loss minimization problems.However, many applications of interest, like minimizing projected Bellman error and min-max optimization, cannot be modelled as minimizing a scalar loss function but instead correspond to solving a variational inequality (VI) problem.This difference in setting has caused many practical challenges as naive gradient-based approaches from supervised learning tend to diverge and cycle in the VI case.In this work, we propose a principled surrogate-based approach compatible with deep learning to solve VIs.We show that our surrogate-based approach has three main benefits: (1) under assumptions that are realistic in practice (when hidden monotone structure is present, interpolation, and sufficient optimization of the surrogates), it guarantees convergence, (2) it provides a unifying perspective of existing methods, and (3) is amenable to existing deep learning optimizers like ADAM.Experimentally, we demonstrate our surrogate-based approach is effective in min-max optimization and minimizing projected Bellman error. Furthermore, in the deep reinforcement learning case, we propose a novel variant of TD(0) which is more compute and sample efficient.

729. HELMET: How to Evaluate Long-context Models Effectively and Thoroughly

链接：https://iclr.cc/virtual/2025/poster/31157 abstract： Many benchmarks exist for evaluating long-context language models (LCLMs), yet developers often rely on synthetic tasks such as needle-in-a-haystack (NIAH) or an arbitrary subset of tasks. However, it remains unclear whether these benchmarks reflect the diverse downstream applications of LCLMs, and such inconsistencies further complicate model comparison. We investigate the underlying reasons behind these practices and find that existing benchmarks often provide noisy signals due to limited coverage of applications, insufficient context lengths, unreliable metrics, and incompatibility with base models. In this work, we introduce HELMET (How to Evaluate Long-context

Models Effectively and Thoroughly), a comprehensive benchmark encompassing seven diverse, application-centric categories. We also address several issues in previous benchmarks by adding controllable lengths up to 128K tokens, model-based evaluation for reliable metrics, and few-shot prompting for robustly evaluating base models. Consequently, we demonstrate that HELMET offers more reliable and consistent rankings of frontier LCLMs. Through a comprehensive study of 59 LCLMs, we find that (1) synthetic tasks like NIAH do not reliably predict downstream performance; (2) the diverse categories in HELMET exhibit distinct trends and low correlations with each other; and (3) while most LCLMs achieve perfect NIAH scores, open-source models significantly lag behind closed ones when tasks require full-context reasoning or following complex instructions---the gap widens as length increases. Finally, we recommend using our RAG tasks for fast model development, as they are easy to run and better predict other downstream performance; ultimately, we advocate for a holistic evaluation across diverse tasks.

730. Universal Sharpness Dynamics in Neural Network Training: Fixed Point Analysis, Edge of Stability, and Route to Chaos

链接：https://iclr.cc/virtual/2025/poster/29406 abstract： In gradient descent dynamics of neural networks, the top eigenvalue of the Hessian of the loss (sharpness) displays a variety of robust phenomena throughout training. This includes early time regimes where the sharpness may decrease during early periods of training (sharpness reduction), and later time behavior such as progressive sharpening and edge of stability. We demonstrate that a simple -layer linear network (UV model) trained on a single training example exhibits all of the essential sharpness phenomenology observed in real-world scenarios. By analyzing the structure of dynamical fixed points in function space and the vector field of function updates, we uncover the underlying mechanisms behind these sharpness trends. Our analysis reveals (i) the mechanism behind early sharpness reduction and progressive sharpening, (ii) the required conditions for edge of stability, and (iii) a period-doubling route to chaos on the edge of stability manifold as learning rate is increased. Finally, we demonstrate that various predictions from this simplified model generalize to real-world scenarios and discuss its limitations.

731. Diffusion-based Neural Network Weights Generation

链接：https://iclr.cc/virtual/2025/poster/28652 abstract： Transfer learning is a cornerstone of modern deep learning, yet it remains constrained by challenges in model selection and the overhead of extensive model storage. In this work, we present Diffusion-based Neural Network Weights Generation, D2NWG, a novel framework that leverages diffusion processes to synthesize task-specific network weights. By modeling the distribution of weights from a diverse ensemble of pretrained models and conditioning the generation process on dataset characteristics, task descriptions, and architectural specifications, D2NWG circumvents the need for storing and searching through massive model repositories. We evaluate D2NWG across multiple experimental settings. On in-distribution tasks, our framework achieves performance that is on par with or superior to conventional pretrained models, while also serving as an effective initialization strategy for novel domains, resulting in faster convergence and a 6\% improvement in few-shot learning scenarios. Extensive ablation studies further indicate that our approach scales robustly with increased diversity and volume of pretrained models. Moreover, D2NWG demonstrates significant promise for large language model applications. In evaluations on the OpenLM leaderboard, our method improved LLaMA-3-2-1B-Instruct performance by 3\% on challenging mathematical reasoning tasks, with a consistent gain of 0.36\% across a range of benchmarks. These findings establish D2NWG as a versatile and powerful framework for neural network weight generation, offering a scalable solution to the limitations of traditional transfer learning.

732. Fitting Networks with a Cancellation Trick

链接：https://iclr.cc/virtual/2025/poster/30535 abstract： The degree-corrected block model (DCBM), latent space model (LSM), and β-model are all popular network models. We combine their modeling ideas and propose the logit-DCBM as a new model. Similar as the β-model and LSM, the logit-DCBM contains nonlinear factors, where fitting the parameters is a challenging open problem. We resolve this problem by introducing a cancellation trick. We also propose R-SCORE as a recursive community detection algorithm, where in each iteration, we first use the idea above to update our parameter estimation, and then use the results to remove the nonlinear factors in the logit-DCBM so the renormalized model approximately satisfies a low-rank model, just like the DCBM. Our numerical study suggests that R-SCORE significantly improves over existing spectral approaches in many cases. Also, theoretically, we show that the Hamming error rate of R-SCORE is faster than that of SCORE in a specific sparse region, and is at least as fast outside this region.

733. Learning Diverse Attacks on Large Language Models for Robust Red-Teaming and Safety Tuning

链接：https://iclr.cc/virtual/2025/poster/31186 abstract： Red-teaming, or identifying prompts that elicit harmful responses, is a critical step in ensuring the safe and responsible deployment of large language models (LLMs). Developing effective protection against many modes of attack prompts requires discovering diverse attacks. Automated red-teaming typically uses reinforcement learning to fine-tune an attacker language model to generate prompts that elicit undesirable responses from a target LLM, as measured, for example, by an auxiliary toxicity classifier. We show that even with explicit regularization to favor novelty and diversity, existing approaches suffer from mode collapse or fail to generate effective attacks. As a flexible and probabilistically principled alternative, we propose to use GFlowNet fine-tuning, followed by a secondary smoothing phase, to train the attacker model to generate diverse and effective attack prompts. We find that the attacks generated by our method are effective against a wide range of target LLMs, both with and without safety tuning, and transfer well between target LLMs. Finally,

we demonstrate that models safety-tuned using a dataset of red-teaming prompts generated by our method are robust to attacks from other RL-based red-teaming approaches.

734. Poisson-Dirac Neural Networks for Modeling Coupled Dynamical Systems across Domains

链接：https://iclr.cc/virtual/2025/poster/29495 abstract： Deep learning has achieved great success in modeling dynamical systems, providing data-driven simulators to predict complex phenomena, even without known governing equations. However, existing models have two major limitations: their narrow focus on mechanical systems and their tendency to treat systems as monolithic. These limitations reduce their applicability to dynamical systems in other domains, such as electrical and hydraulic systems, and to coupled systems. To address these limitations, we propose Poisson-Dirac Neural Networks (PoDiNNs), a novel framework based on the Dirac structure that unifies the port-Hamiltonian and Poisson formulations from geometric mechanics. This framework enables a unified representation of various dynamical systems across multiple domains as well as their interactions and degeneracies arising from couplings. Our experiments demonstrate that PoDiNNs offer improved accuracy and interpretability in modeling unknown coupled dynamical systems from data.

735. AssembleFlow: Rigid Flow Matching with Inertial Frames for Molecular Assembly

链接：https://iclr.cc/virtual/2025/poster/28629 abstract： Molecular assembly, where a cluster of rigid molecules aggregated into strongly correlated forms, is fundamental to determining the properties of materials. However, traditional numerical methods for simulating this process are computationally expensive, and existing generative models on material generation overlook the rigidity inherent in molecular structures, leading to unwanted distortions and invalid internal structures in molecules. To address this, we introduce AssembleFlow. AssembleFlow leverages inertial frames to establish reference coordinate systems at the molecular level for tracking the orientation and motion of molecules within the cluster. It further decomposes molecular $\text{SE}(3)$ transformations into translations in $\mathbb{R}^3$ and rotations in $\text{SO}(3)$, enabling explicit enforcement of both translational and rotational rigidity during each generation step within the flow matching framework. This decomposition also empowers distinct probability paths for each transformation group, effectively allowing for the separate learning of their velocity functions: the former, moving in Euclidean space, uses linear interpolation (LERP), while the latter, evolving in spherical space, employs spherical linear interpolation (SLERP) with a closed-form solution. Empirical validation on the benchmarking data COD-Cluster17 shows that AssembleFlow significantly outperforms six competitive deep learning baselines by at least 45\% in assembly matching scores while maintaining 100\% molecular integrity. Also, it matches the assembly performance of a widely used domain-specific simulation tool while reducing computational cost by 25-fold.

736. Sparse Autoencoders Do Not Find Canonical Units of Analysis

链接：https://iclr.cc/virtual/2025/poster/30676 abstract： A common goal of mechanistic interpretability is to decompose the activations of neural networks into features: interpretable properties of the input computed by the model. Sparse autoencoders (SAEs) are a popular method for finding these features in LLMs, and it has been postulated that they can be used to find a canonical set of units: a unique and complete list of atomic features. We cast doubt on this belief using two novel techniques: SAE stitching to show they are incomplete, and meta-SAEs to show they are not atomic. SAE stitching involves inserting or swapping latents from a larger SAE into a smaller one. Latents from the larger SAE can be divided into two categories: novel latents, which improve performance when added to the smaller SAE, indicating they capture novel information, and reconstruction latents, which can replace corresponding latents in the smaller SAE that have similar behavior. The existence of novel features indicates incompleteness of smaller SAEs. Using meta-SAEs - SAEs trained on the decoder matrix of another SAE - we find that latents in SAEs often decompose into combinations of latents from a smaller SAE, showing that larger SAE latents are not atomic. The resulting decompositions are often interpretable; e.g. a latent representing "Einstein" decomposes into "scientist", "Germany", and "famous person". To train meta-SAEs we introduce BatchTopK SAEs, an improved variant of the popular TopK SAE method, that only enforces a fixed average sparsity. Even if SAEs do not find canonical units of analysis, they may still be useful tools. We suggest that future research should either pursue different approaches for identifying such units, or pragmatically choose the SAE size suited to their task. We provide an interactive dashboard to explore meta-SAEs: https://metasaes.streamlit.app/

737. Selective Unlearning via Representation Erasure Using Domain Adversarial Training

链接：https://iclr.cc/virtual/2025/poster/30023 abstract： When deploying machine learning models in the real world, we often face the challenge of “unlearning” specific data points or subsets after training. Inspired by Domain-Adversarial Training of Neural Networks (DANN), we propose a novel algorithm,SURE, for targeted unlearning.SURE treats the process as a domain adaptation problem, where the “forget set” (data to be removed) and a validation set from the same distribution form two distinct domains. We train a domain classifier to discriminate between representations from the forget and validation sets.Using a gradient reversal strategy similar to DANN, we perform gradient updates to the representations to “fool” the domain classifier and thus obfuscate representations belonging to the forget set. Simultaneously, gradient descent is applied to the retain set (original training data minus the forget set) to preserve its classification performance. Unlike other unlearning approaches whose

training objectives are built based on model outputs, SURE directly manipulates the representations.This is key to ensure robustness against a set of more powerful attacks than currently considered in the literature, that aim to detect which examples were unlearned through access to learned embeddings. Our thorough experiments reveal that SURE has a better unlearning quality to utility trade-off compared to other standard unlearning techniques for deep neural networks.

738. Don't flatten, tokenize! Unlocking the key to SoftMoE's efficacy in deep RL

链接：https://iclr.cc/virtual/2025/poster/30744 abstract： The use of deep neural networks in reinforcement learning (RL) often suffers from performance degradation as model size increases. While soft mixtures of experts (SoftMoEs) have recently shown promise in mitigating this issue for online RL, the reasons behind their effectiveness remain largely unknown. In this work we provide an in-depth analysis identifying the key factors driving this performance gain. We discover the surprising result that tokenizing the encoder output, rather than the use of multiple experts, is what is behind the efficacy of SoftMoEs. Indeed, we demonstrate that even with an appropriately scaled single expert, we are able to maintain the performance gains, largely thanks to tokenization.

739. Dynamic Low-Rank Sparse Adaptation for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28352 abstract： Despite the efficacy of network sparsity in alleviating the deployment strain of Large Language Models (LLMs), it endures significant performance degradation. Applying Low-Rank Adaptation (LoRA) to fine-tune the sparse LLMs offers an intuitive approach to counter this predicament, while it holds shortcomings include: 1) The inability to integrate LoRA weights into sparse LLMs post-training, and 2) Insufficient performance recovery at high sparsity ratios. In this paper, we introduces dynamic $\textbf{Lo}$w-rank $\textbf{S}$parse $\textbf{A}$daptation $\textbf{(LoSA)}$, a novel method that seamlessly integrates low-rank adaptation into LLM sparsity within a unified framework, thereby enhancing the performance of sparse LLMs without increasing the inference latency. In particular, LoSA dynamically sparsifies the LoRA outcomes based on the corresponding sparse weights during fine-tuning, thus guaranteeing that the LoRA module can be integrated into the sparse LLMs post-training. Besides, to achieve the optimal sparse model architecture, LoSA leverages Representation Mutual Information (RMI) as an indicator to determine the importance of layers, thereby dynamically determining the optimal layer-wise sparsity rates during fine-tuning. Predicated on this, LoSA adjusts the rank of the LoRA module based on the variability in layer-wise reconstruction errors, allocating an appropriate fine-tuning for each layer to reduce the output discrepancies between dense and sparse LLMs. Extensive experiments tell that LoSA can efficiently boost the efficacy of sparse LLMs within a few hours, without introducing any additional inferential burden. For example, LoSA reduced the perplexity of sparse LLaMA-2-7B by $\textbf{68.73}$$\downarrow$ and increased zero-shot accuracy by $\textbf{16.32}$%$\uparrow$, achieving a $\textbf{2.60$\times$}$ speedup on CPU and $\textbf{2.23$\times$}$ speedup on GPU, requiring only $\textbf{45 minutes}$ of fine-tuning on $\textbf{a single}$ NVIDIA A100 80GB GPU. Code is available at https://github.com/wzhuang-xmu/LoSA.

740. Learning a Fast Mixing Exogenous Block MDP using a Single Trajectory

链接：https://iclr.cc/virtual/2025/poster/31043 abstract： In order to train agents that can quickly adapt to new objectives or reward functions, efficient unsupervised representation learning in sequential decision-making environments can be important. Frameworks such as the Exogenous Block Markov Decision Process (Ex-BMDP) have been proposed to formalize this representation-learning problem (Efroni et al., 2022b). In the Ex-BMDP framework, the agent's high-dimensional observations of the environment have two latent factors: a controllable factor, which evolves deterministically within a small state space according to the agent's actions, and an exogenous factor, which represents time-correlated noise, and can be highly complex. The goal of the representation learning problem is to learn an encoder that maps from observations into the controllable latent space, as well as the dynamics of this space. Efroni et al. (2022b) has shown that this is possible with a sample complexity that depends only on the size of the controllable latent space, and not on the size of the noise factor. However, this prior work has focused on the episodic setting, where the controllable latent state resets to a specific start state after a finite horizon.By contrast, if the agent can only interact with the environment in a single continuous trajectory, prior works have not established sample-complexity bounds. We propose STEEL, the first provably sample-efficient algorithm for learning the controllable dynamics of an Ex-BMDP from a single trajectory, in the function approximation setting. STEEL has a sample complexity that depends only on the sizes of the controllable latent space and the encoder function class, and (at worst linearly) on the mixing time of the exogenous noise factor. We prove that STEEL is correct and sample-efficient, and demonstrate STEEL on two toy problems. Code is available at: https://github.com/midi-lab/steel.

741. Conflict-Averse Gradient Aggregation for Constrained Multi-Objective Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28337 abstract： In real-world applications, a reinforcement learning (RL) agent should consider multiple objectives and adhere to safety guidelines.To address these considerations, we propose a constrained multi-objective RL algorithm named constrained multi-objective gradient aggregator (CoMOGA).In the field of multi-objective optimization, managing conflicts between the gradients of the multiple objectives is crucial to prevent policies from converging to local optima.It is also essential to efficiently handle safety constraints for stable training and constraint satisfaction.We address these challenges straightforwardly by treating the maximization of multiple objectives as a constrained optimization problem

(COP), where the constraints are defined to improve the original objectives.Existing safety constraints are then integrated into the COP, and the policy is updated by solving the COP, which ensures the avoidance of gradient conflicts.Despite its simplicity, CoMOGA guarantees convergence to global optima in a tabular setting.Through various experiments, we have confirmed that preventing gradient conflicts is critical, and the proposed method achieves constraint satisfaction across all tasks.

742. CAKE: Cascading and Adaptive KV Cache Eviction with Layer Preferences

链接：https://iclr.cc/virtual/2025/poster/30406 abstract： Large language models (LLMs) excel at processing long sequences, boosting demand for key-value (KV) caching. While recent efforts to evict KV cache have alleviated the inference burden, they often fail to allocate resources rationally across layers with different attention patterns. In this paper, we introduce Cascading and Adaptive KV cache Eviction (CAKE), a novel approach that frames KV cache eviction as a ``cake-slicing problem."CAKE assesses layer-specific preferences by considering attention dynamics in both spatial and temporal dimensions, allocates rational cache size for layers accordingly, and manages memory constraints in a cascading manner. This approach enables a global view of cache allocation, adaptively distributing resources across diverse attention mechanisms while maintaining memory budgets.CAKE also employs a new eviction indicator that considers the shifting importance of tokens over time, addressing limitations in existing methods that overlook temporal dynamics.Comprehensive experiments on LongBench and NeedleBench show that CAKE maintains model performance with only 3.2\% of the KV cache and consistently outperforms current baselines across various models and memory constraints, particularly in low-memory settings. Additionally, CAKE achieves over 10$\times$ speedup in decoding latency compared to full cache when processing contexts of 128K tokens with FlashAttention-2. Our code is available at https://github.com/antgroup/cakekv.

743. Re-Thinking Inverse Graphics With Large Language Models

链接：https://iclr.cc/virtual/2025/poster/31472 abstract： Inverse graphics -- the task of inverting an image into physical variables that, when rendered, enable reproduction of the observed scene -- is a fundamental challenge in computer vision and graphics. Successfully disentangling an image into its constituent elements, such as the shape, color, and material properties of the objects of the 3D scene that produced it, requires a comprehensive understanding of the environment. This complexity limits the ability of existing carefully engineered approaches to generalize across domains. Inspired by the zero-shot ability of large language models (LLMs) to generalize to novel contexts, we investigate the possibility of leveraging the broad world knowledge encoded in such models to solve inverse-graphics problems. To this end, we propose the Inverse-Graphics Large Language Model (IG-LLM), an inverse-graphics framework centered around an LLM, that autoregressively decodes a visual embedding into a structured, compositional 3D-scene representation. We incorporate a frozen pre-trained visual encoder and a continuous numeric head to enable end-to-end training. Through our investigation, we demonstrate the potential of LLMs to facilitate inverse graphics through next-token prediction, without the application of image-space supervision. Our analysis enables new possibilities for precise spatial reasoning about images that exploit the visual knowledge of LLMs. We release our code and data at https://ig-llm.is.tue.mpg.de/ to ensure the reproducibility of our investigation and to facilitate future research.

744. Rethinking Self-Distillation: Label Averaging and Enhanced Soft Label Refinement with Partial Labels

链接：https://iclr.cc/virtual/2025/poster/30413 abstract： We investigate the mechanisms of self-distillation in multi-class classification, particularly in the context of linear probing with fixed feature extractors where traditional feature learning explanations do not apply. Our theoretical analysis reveals that multi-round self-distillation effectively performs label averaging among instances with high feature correlations, governed by the eigenvectors of the Gram matrix derived from input features. This process leads to clustered predictions and improved generalization, mitigating the impact of label noise by reducing the model's reliance on potentially corrupted labels. We establish conditions under which multi-round self-distillation achieves 100\% population accuracy despite label noise. Furthermore, we introduce a novel, efficient single-round self-distillation method using refined partial labels from the teacher's top two softmax outputs, referred to as the PLL student model. This approach replicates the benefits of multi-round distillation in a single round, achieving comparable or superior performance--especially in high-noise scenarios--while significantly reducing computational cost.

745. Identifiability for Gaussian Processes with Holomorphic Kernels

链接：https://iclr.cc/virtual/2025/poster/30338 abstract： Gaussian processes (GPs) are widely recognized for their robustness and flexibility across various domains, including machine learning, time series, spatial statistics, and biomedicine. In addition to their common usage in regression tasks, GP kernel parameters are frequently interpreted in various applications. For example, in spatial transcriptomics, estimated kernel parameters are used to identify spatial variable genes, which exhibit significant expression patterns across different tissue locations. However, before these parameters can be meaningfully interpreted, it is essential to establish their identifiability. Existing studies of GP parameter identifiability have focused primarily on Mat\'ern-type kernels, as their spectral densities allow for more established mathematical tools. In many real-world applications, particularly in time series analysis, other kernels such as the squared exponential, periodic, and rational quadratic kernels, as well as their combinations, are also widely used. These kernels share the property of being holomorphic around zero, and their parameter identifiability remains underexplored.In this paper, we bridge this gap by developing a novel theoretical framework for determining kernel parameter identifiability for kernels holomorphic near zero. Our findings enable practitioners to determine

which parameters are identifiable in both existing and newly constructed kernels, supporting application-specific interpretation of the identifiable parameters, and highlighting non-identifiable parameters that require careful interpretation.

746. Neural Dueling Bandits: Preference-Based Optimization with Human Feedback

链接：https://iclr.cc/virtual/2025/poster/29425 abstract： Contextual dueling bandit is used to model the bandit problems, where a learner's goal is to find the best arm for a given context using observed noisy human preference feedback over the selected arms for the past contexts. However, existing algorithms assume the reward function is linear, which can be complex and non-linear in many real-life applications like online recommendations or ranking web search results. To overcome this challenge, we use a neural network to estimate the reward function using preference feedback for the previously selected arms. We propose upper confidence bound- and Thompson sampling-based algorithms with sub-linear regret guarantees that efficiently select arms in each round. We also extend our theoretical results to contextual bandit problems with binary feedback, which is in itself a non-trivial contribution. Experimental results on the problem instances derived from synthetic datasets corroborate our theoretical results.

747. Residual Deep Gaussian Processes on Manifolds

链接：https://iclr.cc/virtual/2025/poster/30105 abstract： We propose practical deep Gaussian process models on Riemannian manifolds, similar in spirit to residual neural networks.With manifold-to-manifold hidden layers and an arbitrary last layer, they can model manifold- and scalar-valued functions, as well as vector fields.We target data inherently supported on manifolds, which is too complex for shallow Gaussian processes thereon.For example, while the latter perform well on high-altitude wind data, they struggle with the more intricate, nonstationary patterns at low altitudes.Our models significantly improve performance in these settings, enhancing prediction quality and uncertainty calibration, and remain robust to overfitting, reverting to shallow models when additional complexity is unneeded.We further showcase our models on Bayesian optimisation problems on manifolds, using stylised examples motivated by robotics, and obtain substantial improvements in later stages of the optimisation process.Finally, we show our models to have potential for speeding up inference for non-manifold data, when, and if, it can be mapped to a proxy manifold well enough.

748. HADAMRNN: BINARY AND SPARSE TERNARY ORTHOGONAL RNNS

链接：https://iclr.cc/virtual/2025/poster/29150 abstract： Binary and sparse ternary weights in neural networks enable faster computations and lighter representations, facilitating their use on edge devices with limited computational power. Meanwhile, vanilla RNNs are highly sensitive to changes in their recurrent weights, making the binarization and ternarization of these weights inherently challenging. To date, no method has successfully achieved binarizationor ternarization of vanilla RNN weights. We present a new approach leveraging the properties of Hadamard matrices to parameterize a subset of binary and sparse ternary orthogonal matrices. This method enables the training of orthogonal RNNs (ORNNs) with binary and sparse ternary recurrent weights, effectively creating a specific class of binary and sparse ternary vanilla RNNs. The resulting ORNNs, called HadamRNN and Block-HadamRNN, are evaluated on benchmarks such as the copy task, permuted and sequential MNIST tasks, and IMDB dataset. Despite binarization or sparse ternarization, these RNNs maintain performance levels comparable to state-of-the-art full-precision models, highlighting the effectiveness of our approach. Notably, our approach is the first solution with binary recurrent weights capable of tackling the copy task over 1000 timesteps.

749. U-Nets as Belief Propagation: Efficient Classification, Denoising, and Diffusion in Generative Hierarchical Models

链接：https://iclr.cc/virtual/2025/poster/28085 abstract： U-Nets are among the most widely used architectures in computer vision, renowned for their exceptional performance in applications such as image segmentation, denoising, and diffusion modeling. However, a theoretical explanation of the U-Net architecture design has not yet been fully established. This paper introduces a novel interpretation of the U-Net architecture by studying certain generative hierarchical models, which are tree-structured graphical models extensively utilized in both language and image domains. With their encoder-decoder structure, long skip connections, and pooling and up-sampling layers, we demonstrate how U-Nets can naturally implement the belief propagation denoising algorithm in such generative hierarchical models, thereby efficiently approximating the denoising functions. This leads to an efficient sample complexity bound for learning the denoising function using U-Nets within these models. Additionally, we discuss the broader implications of these findings for diffusion models in generative hierarchical models. We also demonstrate that the conventional architecture of convolutional neural networks (ConvNets) is ideally suited for classification tasks within these models. This offers a unified view of the roles of ConvNets and U-Nets, highlighting the versatility of generative hierarchical models in modeling complex data distributions.

750. Probabilistic Language-Image Pre-Training

链接：https://iclr.cc/virtual/2025/poster/30478 abstract： Vision-language models (VLMs) embed aligned image-text pairs into a joint space but often rely on deterministic embeddings, assuming a one-to-one correspondence between images and texts. This oversimplifies real-world relationships, which are inherently many-to-many, with multiple captions describing a single image

and vice versa. We introduce Probabilistic Language-Image Pre-training (ProLIP), the first probabilistic VLM pre-trained on a billion-scale image-text dataset using only probabilistic objectives, achieving a strong zero-shot capability (e.g., 74.6% ImageNet zero-shot accuracy with ViT-B/16). ProLIP efficiently estimates uncertainty by an ``uncertainty token'' without extra parameters. We also introduce a novel inclusion loss that enforces distributional inclusion relationships between image-text pairs and between original and masked inputs. Experiments demonstrate that, by leveraging uncertainty estimates, ProLIP benefits downstream tasks and aligns with intuitive notions of uncertainty, e.g., shorter texts being more uncertain and more general inputs including specific ones. Utilizing text uncertainties, we further improve ImageNet accuracy from 74.6% to 75.8% (under a few-shot setting), supporting the practical advantages of our probabilistic approach. The code is available at https://github.com/naver-ai/prolip

751. ϕ-Update: A Class of Policy Update Methods with Policy Convergence Guarantee

链接：https://iclr.cc/virtual/2025/poster/28864 abstract： Inspired by the similar update pattern of softmax natural policy gradient and Hadamard policy gradient, we propose to study a general policy update rule called ϕ-update, where ϕ refers to a scaling function on advantage functions. Under very mild conditions on ϕ, the global asymptotic state value convergence of ϕ-update is firstly established. Then we show that the policy produced by ϕ-update indeed converges, even when there are multiple optimal policies. This is in stark contrast to existing results where explicit regularizations are required to guarantee the convergence of the policy. Since softmax natural policy gradient is an instance of ϕ-update, it provides an affirmative answer to the question whether the policy produced by softmax natural policy gradient converges. The exact asymptotic convergence rate of state values is further established based on the policy convergence. Lastly, we establish the global linear convergence of ϕ-update.

752. NetFormer: An interpretable model for recovering dynamical connectivity in neuronal population dynamics

链接：https://iclr.cc/virtual/2025/poster/29102 abstract： Neuronal dynamics are highly nonlinear and nonstationary. Traditional methods for extracting the underlying network structure from neuronal activity recordings mainly concentrate on modeling static connectivity, without accounting for key nonstationary aspects of biological neural systems, such as ongoing synaptic plasticity and neuronal modulation. To bridge this gap, we introduce the NetFormer model, an interpretable approach applicable to such systems. In NetFormer, the activity of each neuron across a series of historical time steps is defined as a token. These tokens are then linearly mapped through a query and key mechanism to generate a state- (and hence time-) dependent attention matrix that directly encodes nonstationary connectivity structures. We analyze our formulation from the perspective of nonstationary and nonlinear networked dynamical systems, and show both via an analytical expansion and targeted simulations how it can approximate the underlying ground truth. Next, we demonstrate NetFormer's ability to model a key feature of biological networks, spike-timing-dependent plasticity, whereby connection strengths continually change in response to local activity patterns. We further demonstrate that NetFormer can capture task-induced connectivity patterns on activity generated by task-trained recurrent neural networks. Thus informed, we apply NetFormer to a multi-modal dataset of real neural recordings, which contains neural activity, cell type, and behavioral state information. We show that the NetFormer effectively predicts neural dynamics and identifies cell-type specific, state-dependent dynamic connectivity that matches patterns measured in separate ground-truth physiology experiments, demonstrating its ability to help decode complex neural interactions based on population activity observations alone.

753. PaPaGei: Open Foundation Models for Optical Physiological Signals

链接：https://iclr.cc/virtual/2025/poster/28573 abstract： Photoplethysmography (PPG) is the leading non-invasive technique for monitoring biosignals and cardiovascular health, with widespread adoption in both clinical settings and consumer wearable devices. While machine learning models trained on PPG signals have shown promise, they tend to be task-specific and struggle with generalization. Current research is limited by the use of single-device datasets, insufficient exploration of out-of-domain generalization, and a lack of publicly available models, which hampers reproducibility. To address these limitations, we present PaPaGei, the first open foundation model for PPG signals. The model is pre-trained on over 57,000 hours of data, comprising 20 million unlabeled PPG segments from publicly available datasets. We introduce a novel representation learning approach that leverages domain knowledge of PPG signal morphology across individuals, enabling the capture of richer representations compared to traditional contrastive learning methods. We evaluate PaPaGei against state-of-the-art time-series foundation models and self-supervised learning benchmarks across 20 tasks from 10 diverse datasets, spanning cardiovascular health, sleep disorders, pregnancy monitoring, and wellbeing assessment. Our model demonstrates superior performance, improving classification and regression metrics by 6.3% and 2.9% respectively in at least 14 tasks. Notably, PaPaGei achieves these results while being more data- and parameter-efficient, outperforming models that are 70x larger. Beyond accuracy, we examine model robustness across different skin tones, establishing a benchmark for bias evaluation in future models. PaPaGei can serve as both a feature extractor and an encoder for multimodal models, opening up new opportunities for multimodal health monitoring. Models, data, and code are available at: https://github.com/nokia-bell-labs/papagei-foundation-model

754. Correcting the Mythos of KL-Regularization: Direct Alignment without Overoptimization via Chi-Squared Preference Optimization

链接：https://iclr.cc/virtual/2025/poster/28754 abstract： Language model alignment methods such as reinforcement learning from human feedback (RLHF) have led to impressive advances in language model capabilities, but are limited by a widely observed phenomenon known as *overoptimization*, where the quality of the language model degrades over the course of the alignment process. As the model optimizes performance on an offline reward model, it overfits to inaccuracies and drifts away from preferred responses covered by the data. To discourage such distribution shift, KL-regularization is widely employed in existing offline alignment methods, but overoptimization continues to harm performance. Lending theoretical insight into the source of these empirical observations, we first show that the KL-regularization is too weak to prevent overfitting, then ask: is it possible to design an efficient algorithm that is provably robust to overoptimization?In this paper, we advance theoretical understanding of sample-efficient offline alignment and introduce a new algorithm called χ^2-Preference Optimization (χPO). χPO is a one-line change to Direct Preference Optimization (DPO; Rafailov et al. 2023), that modifies only the logarithmic link function in the DPO objective. Despite this minimal change, χPO implicitly implements the principle of *pessimism in the face of uncertainty* via regularization with the χ^2-divergence---which quantifies uncertainty more effectively than KL-regularization---and provably alleviates overoptimization, achieving sample-complexity guarantees based on *single-policy concentrability*, the gold standard in offline reinforcement learning. This guarantee makes χPO the first simple, yet general-purpose offline alignment algorithm that is provably robust to overoptimization.

755. Towards Improving Exploration through Sibling Augmented GFlowNets

链接：https://iclr.cc/virtual/2025/poster/30233 abstract： Exploration is a key factor for the success of an active learning agent, especially when dealing with sparse extrinsic terminal rewards and long trajectories. We introduce Sibling Augmented Generative Flow Networks (SA-GFN), a novel framework designed to enhance exploration and training efficiency of Generative Flow Networks (GFlowNets). SA-GFN uses a decoupled dual network architecture, comprising of a main Behavior Network and an exploratory Sibling Network, to enable a diverse exploration of the underlying distribution using intrinsic rewards. Inspired by the ideas on exploration from reinforcement learning, SA-GFN provides a general-purpose exploration and learning paradigm that integrates with multiple GFlowNet training objectives and is especially helpful for exploration over a wide range of sparse or low reward distributions and task structures. An extensive set of experiments across a diverse range of tasks, reward structures and trajectory lengths, along with a thorough set of ablations, demonstrate the superior performance of SA-GFN in terms of exploration efficacy and convergence speed as compared to the existing methods. In addition, SA-GFN's versatility and compatibility with different GFlowNet training objectives and intrinsic reward methods underscores its broad applicability in various problem domains.

756. PianoMotion10M: Dataset and Benchmark for Hand Motion Generation in Piano Performance

链接：https://iclr.cc/virtual/2025/poster/28158 abstract： Recently, artificial intelligence techniques for education have been received increasing attentions, while it still remains an open problem to design the effective music instrument instructing systems. Although key presses can be directly derived from sheet music, the transitional movements among key presses require more extensive guidance in piano performance. In this work, we construct a piano-hand motion generation benchmark to guide hand movements and fingerings for piano playing. To this end, we collect an annotated dataset, PianoMotion10M, consisting of 116 hours of piano playing videos from a bird's-eye view with 10 million annotated hand poses. We also introduce a powerful baseline model that generates hand motions from piano audios through a position predictor and a position-guided gesture generator. Furthermore, a series of evaluation metrics are designed to assess the performance of the baseline model, including motion similarity, smoothness, positional accuracy of left and right hands, and overall fidelity of movement distribution. Despite that piano key presses with respect to music scores or audios are already accessible, PianoMotion10M aims to provide guidance on piano fingering for instruction purposes. The source code and dataset can be accessed at https://github.com/agnJason/PianoMotion10M.

757. On the Transfer of Object-Centric Representation Learning

链接：https://iclr.cc/virtual/2025/poster/29108 abstract： The goal of object-centric representation learning is to decompose visual scenes into a structured representation that isolates the entities into individual vectors. Recent successes have shown that object-centric representation learning can be scaled to real-world scenes by utilizing features from pre-trained foundation models like DINO. However, so far, these object-centric methods have mostly been applied in-distribution, with models trained and evaluated on the same dataset. This is in contrast to the underlying foundation models, which have been shown to be applicable to a wide range of data and tasks. Thus, in this work, we answer the question of whether current real-world capable object-centric methods exhibit similar levels of transferability by introducing a benchmark comprising seven different synthetic and real-world datasets. We analyze the factors influencing performance under transfer and find that training on diverse real-world images improves generalization to unseen scenarios. Furthermore, inspired by the success of task-specific fine-tuning in foundation models, we introduce a novel fine-tuning strategy to adapt pre-trained vision encoders for the task of object discovery. We find that the proposed approach results in state-of-the-art performance for unsupervised object discovery, exhibiting strong zero-shot transfer to unseen datasets.

758. Langevin Soft Actor-Critic: Efficient Exploration through Uncertainty-Driven Critic Learning

链接：https://iclr.cc/virtual/2025/poster/30315 abstract： Existing actor-critic algorithms, which are popular for continuous control reinforcement learning (RL) tasks, suffer from poor sample efficiency due to lack of principled exploration mechanism within them. Motivated by the success of Thompson sampling for efficient exploration in RL, we propose a novel model-free RL algorithm, \emph{Langevin Soft Actor Critic} (LSAC), which prioritizes enhancing critic learning through uncertainty estimation over policy optimization. LSAC employs three key innovations: approximate Thompson sampling through distributional Langevin Monte Carlo (LMC) based Q updates, parallel tempering for exploring multiple modes of the posterior of the Q function, and diffusion synthesized state-action samples regularized with Q action gradients. Our extensive experiments demonstrate that LSAC outperforms or matches the performance of mainstream model-free RL algorithms for continuous control tasks.Notably, LSAC marks the first successful application of an LMC based Thompson sampling in continuous control tasks with continuous action spaces.

759. Structure Language Models for Protein Conformation Generation

链接：https://iclr.cc/virtual/2025/poster/29785 abstract： Proteins adopt multiple structural conformations to perform their diverse biological functions, and understanding these conformations is crucial for advancing drug discovery. Traditional physics-based simulation methods often struggle with sampling equilibrium conformations and are computationally expensive. Recently, deep generative models have shown promise in generating protein conformations as a more efficient alternative. However, these methods predominantly rely on the diffusion process within a 3D geometric space, which typically centers around the vicinity of metastable states and is often inefficient in terms of runtime. In this paper, we introduce Structure Language Modeling (SLM) as a novel framework for efficient protein conformation generation. Specifically, the protein structures are first encoded into a compact latent space using a discrete variational auto-encoder, followed by conditional language modeling that effectively captures sequence-specific conformation distributions. This enables a more efficient and interpretable exploration of diverse ensemble modes compared to existing methods. Based on this general framework, we instantiate SLM with various popular LM architectures as well as proposing the ESMDiff, a novel BERT-like structure language model fine-tuned from ESM3 with masked diffusion. We verify our approach in various scenarios, including the equilibrium dynamics of BPTI, conformational change pairs, and intrinsically disordered proteins. SLM provides a highly efficient solution, offering a 20-100x speedup than existing methods in generating diverse conformations, shedding light on promising avenues for future research.

760. Meta Flow Matching: Integrating Vector Fields on the Wasserstein Manifold

链接：https://iclr.cc/virtual/2025/poster/30690 abstract： Numerous biological and physical processes can be modeled as systems of interacting entities evolving continuously over time, e.g. the dynamics of communicating cells or physical particles. Learning the dynamics of such systems is essential for predicting the temporal evolution of populations across novel samples and unseen environments. Flow-based models allow for learning these dynamics at the population level - they model the evolution of the entire distribution of samples. However, current flow-based models are limited to a single initial population and a set of predefined conditions which describe different dynamics. We argue that multiple processes in natural sciences have to be represented as vector fields on the Wasserstein manifold of probability densities. That is, the change of the population at any moment in time depends on the population itself due to the interactions between samples. In particular, this is crucial for personalized medicine where the development of diseases and their respective treatment response depend on the microenvironment of cells specific to each patient. We propose Meta Flow Matching (MFM), a practical approach to integrate along these vector fields on the Wasserstein manifold by amortizing the flow model over the initial populations. Namely, we embed the population of samples using a Graph Neural Network (GNN) and use these embeddings to train a Flow Matching model. This gives MFM the ability to generalize over the initial distributions, unlike previously proposed methods. We demonstrate the ability of MFM to improve the prediction of individual treatment responses on a large-scale multi-patient single-cell drug screen dataset.

761. How efficient is LLM-generated code? A rigorous & high-standard benchmark

链接：https://iclr.cc/virtual/2025/poster/28090 abstract： The emergence of large language models (LLMs) has significantly pushed the frontiers of program synthesis. Advancement of LLM-based program synthesis calls for a thorough evaluation of LLM-generated code. Most evaluation frameworks focus on the (functional) correctness of generated code; efficiency, as an important measure of code quality, has been overlooked in existing evaluations. In this work, we develop ENAMEL (EfficeNcy AutoMatic EvaLuator), a rigorous and high-standard benchmark for evaluating the capability of LLMs in generating efficient code. Firstly, we propose a new efficiency metric called eff@k, which generalizes the pass@k metric from correctness to efficiency and appropriately handles right-censored execution time. Furthermore, we derive an unbiased and variance-reduced estimator of eff@k via Rao–Blackwellization; we also provide a numerically stable implementation for the new estimator. Secondly, to set a high-standard for efficiency evaluation, we employ a human expert to design best algorithms and implementations as our reference solutions of efficiency, many of which are much more efficient than existing canonical solutions in HumanEval and HumanEval+. Moreover, to ensure a rigorous evaluation, we employ a human expert to curate strong test case generators to filter out wrong code and differentiate suboptimal algorithms. An extensive study across 30 popular LLMs using our benchmark ENAMEL shows that LLMs still fall short of generating expert-level efficient code. Using two subsets of our problem set, we demonstrate that such deficiency is because current LLMs struggle in designing advanced algorithms and are barely aware of implementation optimization.

762. Bootstrapping Language Models with DPO Implicit Rewards

链接：https://iclr.cc/virtual/2025/poster/28967 abstract： Human alignment in large language models (LLMs) is an active area of research. A recent groundbreaking work, direct preference optimization (DPO), has greatly simplified the process from past work in reinforcement learning from human feedback (RLHF) by bypassing the reward learning stage in RLHF. DPO, after training, provides an implicit reward model. In this work, we make a novel observation that this implicit reward model can by itself be used in a bootstrapping fashion to further align the LLM. Our approach is to use the rewards from a current LLM to construct a preference dataset, which is then used in subsequent DPO rounds. We incorporate two refinements to further improve our approach: 1) length-regularized reward shaping to make the preference dataset length-unbiased; 2) experience replay to enhance the quality of the preference dataset. Our approach, named self-alignment with DPO ImpliCit rEwards (DICE), shows great improvements in alignment. It achieves an increase of more than 8$\%$ in lengthcontrolled win rate on AlpacaEval 2 for all the different base models that we tried, without relying on external feedback. Our code is available at https://github.com/sail-sg/dice.

763. Hessian Free Efficient Single Loop Iterative Differentiation Methods for Bi-Level Optimization Problems

链接：https://iclr.cc/virtual/2025/poster/31463 abstract： Bilevel optimization problems have been actively studied in recent machine learning research due to their broad applications. In this work, we investigate single-loop methods with iterative differentiation (ITD) for nonconvex bilevel optimization problems. For deterministic bilevel problems, we propose an efficient single-loop ITD-type method (ES-ITDM). Our method employs historical updates to approximate the hypergradient. More importantly, based on ES-ITDM, we propose a new method that avoids computing Hessians. This Hessian-free method requires fewer backpropagations and thus has a lower computational cost. We analyze the convergence properties of the proposed methods in two aspects. We provide the convergence rates of the sequences generated by ES-ITD based on the Kurdyka-\L ojasiewicz (KL) property. We also show that the Hessian-free stochastic ES-ITDM has the best-known complexity while has cheaper computation. The empirical studies show that our Hessian-free stochastic variant is more efficient than existing Hessian-free methods and other state-of-the-art bilevel optimization approaches.

764. Mix-CPT: A Domain Adaptation Framework via Decoupling Knowledge Learning and Format Alignment

链接：https://iclr.cc/virtual/2025/poster/28784 abstract： Adapting large language models (LLMs) to specialized domains typically requires domain-specific corpora for continual pre-training to facilitate knowledge memorization and related instructions for fine-tuning to apply this knowledge.However, this method may lead to inefficient knowledge memorization due to a lack of awareness of knowledge utilization during the continual pre-training and demands LLMs to simultaneously learn knowledge utilization and format alignment with divergent training objectives during the fine-tuning.To enhance the domain adaptation of LLMs, we revise this process and propose a new domain adaptation framework including domain knowledge learning and general format alignment, called \emph{Mix-CPT}. Specifically, we first conduct a knowledge mixture continual pre-training that concurrently focuses on knowledge memorization and utilization. To avoid catastrophic forgetting, we further propose a logit swap self-distillation constraint. By leveraging the knowledge and capabilities acquired during continual pre-training, we then efficiently perform instruction tuning and alignment with a few general training samples to achieve format alignment.Extensive experiments show that our proposed \emph{Mix-CPT} framework can simultaneously improve the task-solving capabilities of LLMs on the target and general domains.

765. Is Your Video Language Model a Reliable Judge?

链接：https://iclr.cc/virtual/2025/poster/28480 abstract： As video language models (VLMs) gain more applications in various scenarios,the need for robust and scalable evaluation of their performance becomes increasingly critical. The traditional human expert-based evaluation of VLMs has limitations in consistency and scalability, which sparked interest in automatic methodssuch as employing VLMs to evaluate VLMs. However, the reliability of VLMs asjudges remains underexplored. Existing methods often rely on a single VLM asthe evaluator. However, this approach can be unreliable or biased because such amodel may lack the ability to fully understand the content and may have inherentbiases, ultimately compromising evaluation reliability. A remedy is to apply theprinciple of collective thoughts, aggregating evaluations from multiple VLMs toenhance reliability. This study investigates the efficacy of such approaches, particularly when the pool of judges includes both reliable and unreliable models. Ourfindings reveal that incorporating collective judgments from such a mixed pooldoes not necessarily improve the accuracy of the final evaluation. The inclusion ofless reliable judges can introduce noise, undermining the overall reliability of theoutcomes. To explore the factors that impact evaluation reliability, we fine-tunean underperforming VLM judge, Video-LLaVA, and observe that improved understanding ability alone is insufficient to make VLM judges more reliable. Thesefindings stress the limitations of collective thought approaches and highlight theneed for more advanced methods that can account for the reliability of individualmodels. Our study promotes the development of more reliable evaluation methodsfor VLMs

766. A Riemannian Framework for Learning Reduced-order Lagrangian Dynamics

链接：https://iclr.cc/virtual/2025/poster/29638 abstract： By incorporating physical consistency as inductive bias, deep neural networks display increased generalization capabilities and data efficiency in learning nonlinear dynamic models. However, the complexity of these models generally increases with the system dimensionality, requiring larger datasets, more complex deep networks, and significant computational effort.We propose a novel geometric network architecture to learn physically-consistent reduced-order dynamic parameters that accurately describe the original high-dimensional system behavior.This is achieved by building on recent advances in model-order reduction and by adopting a Riemannian perspective to jointly learn a non-linear structure-preserving latent space and the associated low-dimensional dynamics.Our approach enables accurate long-term predictions of the high-dimensional dynamics of rigid and deformable systems with increased data efficiency by inferring interpretable and physically-plausible reduced Lagrangian models.

767. Commit0: Library Generation from Scratch

链接：https://iclr.cc/virtual/2025/poster/29935 abstract： With the goal of benchmarking generative systems beyond expert software development ability, we introduce Commit0, a benchmark that challenges AI agents to write libraries from scratch. Agents are provided with a specification document outlining the library's API as well as a suite of interactive unit tests, with the goal of producing an implementation of this API accordingly. The implementation is validated through running these unit tests. As a benchmark, Commit0 is designed to move beyond static one-shot code generation towards agents that must process long-form natural language specifications, adapt to multi-stage feedback, and generate code with complex dependencies. Commit0 also offers an interactive environment where models receive static analysis and execution feedback on the code they generate. Our experiments demonstrate that while current agents can pass some unit tests, none can yet fully reproduce full libraries. Results also show that interactive feedback is quite useful for models to generate code that passes more unit tests, validating the benchmarks that facilitate its use. We publicly release the benchmark, the interactive environment, and the leaderboard.

768. CraftRTL: High-quality Synthetic Data Generation for Verilog Code Models with Correct-by-Construction Non-Textual Representations and Targeted Code Repair

链接：https://iclr.cc/virtual/2025/poster/30772 abstract： Despite the significant progress made in code generation with large language models, challenges persist, especially with hardware description languages such as Verilog. This paper first presents an analysis of fine-tuned LLMs on Verilog coding, with synthetic data from prior methods. We identify two main issues: difficulties in handling non-textual representations (Karnaugh maps, state-transition diagrams and waveforms) and significant variability during training with models randomly making "minor" mistakes. To address these limitations, we enhance data curation by creating correct-by-construction data targeting non-textual representations. Additionally, we introduce an automated framework that generates error reports from various model checkpoints and injects these errors into open-source code to create targeted code repair data. Our fine-tuned Starcoder2-15B outperforms prior state-of-the-art results by 3.8\%, 10.9\%, 6.6\% for pass@1 on VerilogEval-Machine, VerilogEval-Human, and RTLLM.

769. Latent Bayesian Optimization via Autoregressive Normalizing Flows

链接：https://iclr.cc/virtual/2025/poster/29231 abstract： Bayesian Optimization (BO) has been recognized for its effectiveness in optimizing expensive and complex objective functions.Recent advancements in Latent Bayesian Optimization (LBO) have shown promise by integrating generative models such as variational autoencoders (VAEs) to manage the complexity of high-dimensional and structured data spaces.However, existing LBO approaches often suffer from the value discrepancy problem, which arises from the reconstruction gap between input and latent spaces.This value discrepancy problem propagates errors throughout the optimization process, leading to suboptimal outcomes.To address this issue, we propose a Normalizing Flow-based Bayesian Optimization (NF-BO), which utilizes normalizing flow as a generative model to establish one-to-one encoding function from the input space to the latent space, along with its left-inverse decoding function, eliminating the reconstruction gap. Specifically, we introduce SeqFlow, an autoregressive normalizing flow for sequence data.In addition, we develop a new candidate sampling strategy that dynamically adjusts the exploration probability for each token based on its importance.Through extensive experiments, our NF-BO method demonstrates superior performance in molecule generation tasks, significantly outperforming both traditional and recent LBO approaches.

770. Locality Sensitive Avatars From Video

链接：https://iclr.cc/virtual/2025/poster/29598 abstract： We present locality-sensitive avatar, a neural radiance field (NeRF) based network to learn human motions from monocular videos. To this end, we estimate a canonical representation between different frames of a video with a non-linear mapping from observation to canonical space, which we decompose into a skeletal rigid motion and a non-rigid counterpart. Our key contribution is to retain fine-grained details by modeling the non-rigid part with a graph neural network (GNN) that keeps the pose information local to neighboring body parts. Compared to former canonical representation based methods which solely operate on the coordinate space of a whole shape, our locality-sensitive motion modeling can reproduce both realistic shape contours and vivid fine-grained details. We evaluate on ZJU-MoCap, SynWild, ActorsHQ, MVHumanNet and various outdoor videos. The experiments reveal that with the locality sensitive deformation to canonical feature space, we are the first to achieve state-of-the-art results across novel view synthesis, novel pose animation and 3D shape reconstruction simultaneously. Our code is available at https://github.com/ChunjinSong/lsavatar.

771. Conditional Diffusion with Ordinal Regression: Longitudinal Data Generation for Neurodegenerative Disease Studies

链接：https://iclr.cc/virtual/2025/poster/30688 abstract： Modeling the progression of neurodegenerative diseases such as Alzheimer's disease (AD) is crucial for early detection and prevention given their irreversible nature. However, the scarcity of longitudinal data and complex disease dynamics make the analysis highly challenging. Moreover, longitudinal samples often contain irregular and large intervals between subject visits, which underscore the necessity for advanced data generation techniques that can accurately simulate disease progression over time. In this regime, we propose a novel conditional generative model for synthesizing longitudinal sequences and present its application to neurodegenerative disease data generation conditioned on multiple time-dependent ordinal factors, such as age and disease severity. Our method sequentially generates continuous data by bridging gaps between sparse data points with a diffusion model, ensuring a realistic representation of disease progression. The synthetic data are curated to integrate both cohort-level and individual-specific characteristics, where the cohort-level representations are modeled with an ordinal regression to capture longitudinally monotonic behavior. Extensive experiments on four AD biomarkers validate the superiority of our method over nine baseline approaches, highlighting its potential to be applied to a variety of longitudinal data generation.

772. Action abstractions for amortized sampling

链接：https://iclr.cc/virtual/2025/poster/28667 abstract： As trajectories sampled by policies used by reinforcement learning (RL) and generative flow networks (GFlowNets) grow longer, credit assignment and exploration become more challenging, and the long planning horizon hinders mode discovery and generalization.The challenge is particularly pronounced in entropy-seeking RL methods, such as generative flow networks, where the agent must learn to sample from a structured distribution and discover multiple high-reward states, each of which take many steps to reach.To tackle this challenge, we propose an approach to incorporate the discovery of action abstractions, or high-level actions, into the policy optimization process.Our approach involves iteratively extracting action subsequences commonly used across many high-reward trajectories and `chunking' them into a single action that is added to the action space.In empirical evaluation on synthetic and real-world environments, our approach demonstrates improved sample efficiency performance in discovering diverse high-reward objects, especially on harder exploration problems.We also observe that the abstracted high-order actions are potentially interpretable, capturing the latent structure of the reward landscape of the action space.This work provides a cognitively motivated approach to action abstraction in RL and is the first demonstration of hierarchical planning in amortized sequential sampling.

773. Adaptive teachers for amortized samplers

链接：https://iclr.cc/virtual/2025/poster/30561 abstract： Amortized inference is the task of training a parametric model, such as a neural network, to approximate a distribution with a given unnormalized density where exact sampling is intractable. When sampling is modeled as a sequential decision-making process, reinforcement learning (RL) methods, such as generative flow networks, can be used to train the sampling policy. Off-policy RL training facilitates the discovery of diverse, high-reward candidates, but existing methods still face challenges in efficient exploration. We propose to use an adaptive training distribution (the Teacher) to guide the training of the primary amortized sampler (the Student). The Teacher, an auxiliary behavior model, is trained to sample high-loss regions of the Student and can generalize across unexplored modes, thereby enhancing mode coverage by providing an efficient training curriculum. We validate the effectiveness of this approach in a synthetic environment designed to present an exploration challenge, two diffusion-based sampling tasks, and four biochemical discovery tasks demonstrating its ability to improve sample efficiency and mode coverage. Source code is available at https://github.com/alstn12088/adaptive-teacher.

774. Vision-RWKV: Efficient and Scalable Visual Perception with RWKV-Like Architectures

链接：https://iclr.cc/virtual/2025/poster/28412 abstract： Transformers have revolutionized computer vision and natural language processing, but their high computational complexity limits their application in high-resolution image processing and long-context analysis. This paper introduces Vision-RWKV (VRWKV), a model that builds upon the RWKV architecture from the NLP field with key modifications tailored specifically for vision tasks. Similar to the Vision Transformer (ViT), our model demonstrates robust global processing capabilities, efficiently handles sparse inputs like masked images, and can scale up to accommodate both large-scale parameters and extensive datasets. Its distinctive advantage is its reduced spatial aggregation complexity, enabling seamless processing of high-resolution images without the need for window operations. Our evaluations demonstrate that VRWKV surpasses ViT's performance in image classification and has significantly faster speeds and lower memory usage processing high-resolution inputs. In dense prediction tasks, it outperforms window-based models, maintaining comparable speeds. These results highlight VRWKV's potential as a more efficient alternative for visual perception tasks. Code and models are available at~\url{https://github.com/OpenGVLab/Vision-RWKV}.

775. U-shaped and Inverted-U Scaling behind Emergent Abilities of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28623 abstract： Large language models (LLMs) have been shown to exhibit emergent

abilities in some downstream tasks, where model performance stagnates at first and then improves sharply and unpredictably with scale beyond a threshold. In this work, we investigate the phenomenon by grouping questions based on difficulty level and provide a possible explanation for emergent abilities. Specifically, we observe U-shaped scaling for hard questions and inverted-U scaling followed by steady improvement for easy questions. The two scaling patterns initially offset each other, causing stagnant overall performance. The performance starts to soar when the scaling pattern of easy questions reverts from inverse to standard scaling, leading to emergent abilities. Based on this finding, we propose a simple yet effective pipeline, called Slice-and-Sandwich, to predict the emergence threshold and model performance beyond the threshold. Our code is publicly available at https://github.com/tony10101105/ExpEmergence.

776. Is uniform expressivity too restrictive? Towards efficient expressivity of GNNs

链接：https://iclr.cc/virtual/2025/poster/28495 abstract： Uniform expressivity guarantees that a Graph Neural Network (GNN) can express a query without the parameters depending on the size of the input graphs. This property is desirable in applications in order to have number of trainable parameters that is independent of the size of the input graphs. Uniform expressivity of the two variable guarded fragment (GC2) of first order logic is a well-celebrated result for Rectified Linear Unit (ReLU) GNNs [Barcelo &. Al, 2020]. In this article, we prove that uniform expressivity of GC2 queries is not possible for GNNs with a wide class of Pfaffian activation functions (including the sigmoid and $\tanh$), answering a question formulated by [Grohe, 2021]. We also show that despite these limitations, many of those GNNs can still efficiently express GC2 queries in a way that the number of parameters remains logarithmic on the maximal degree of the input graphs. Furthermore, we demonstrate that a log-log dependency on the degree is achievable for a certain choice of activation function. This shows that uniform expressivity can be successfully relaxed by covering large graphs appearing in practical applications. Our experiments illustrates that our theoretical estimates hold in practice.

777. Tuning Frequency Bias of State Space Models

链接：https://iclr.cc/virtual/2025/poster/27827 abstract： State space models (SSMs) leverage linear, time-invariant (LTI) systems to effectively learn sequences with long-range dependencies. By analyzing the transfer functions of LTI systems, we find that SSMs exhibit an implicit bias toward capturing low-frequency components more effectively than high-frequency ones. This behavior aligns with the broader notion of frequency bias in deep learning model training. We show that the initialization of an SSM assigns it an innate frequency bias and that training the model in a conventional way does not alter this bias. Based on our theory, we propose two mechanisms to tune frequency bias: either by scaling the initialization to tune the inborn frequency bias; or by applying a Sobolev-norm-based filter to adjust the sensitivity of the gradients to high-frequency inputs, which allows us to change the frequency bias via training. Using an image-denoising task, we empirically show that we can strengthen, weaken, or even reverse the frequency bias using both mechanisms. By tuning the frequency bias, we can also improve SSMs' performance on learning long-range sequences, averaging an 88.26% accuracy on the Long-Range Arena (LRA) benchmark tasks.

778. Denoising Task Difficulty-based Curriculum for Training Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/30723 abstract： Diffusion-based generative models have emerged as powerful tools in the realm of generative modeling. Despite extensive research on denoising across various timesteps and noise levels, a conflict persists regarding the relative difficulties of the denoising tasks. While various studies argue that lower timesteps present more challenging tasks, others contend that higher timesteps are more difficult. To address this conflict, our study undertakes a comprehensive examination of task difficulties, focusing on convergence behavior and changes in relative entropy between consecutive probability distributions across timesteps. Our observational study reveals that denoising at earlier timesteps poses challenges characterized by slower convergence and higher relative entropy, indicating increased task difficulty at these lower timesteps. Building on these observations, we introduce an easy-to-hard learning scheme, drawing from curriculum learning, to enhance the training process of diffusion models. By organizing timesteps or noise levels into clusters and training models with ascending orders of difficulty, we facilitate an order-aware training regime, progressing from easier to harder denoising tasks, thereby deviating from the conventional approach of training diffusion models simultaneously across all timesteps. Our approach leads to improved performance and faster convergence by leveraging benefits of curriculum learning, while maintaining orthogonality with existing improvements in diffusion training techniques. We validate these advantages through comprehensive experiments in image generation tasks, including unconditional, class-conditional, and text-to-image generation.

779. Interpreting the Second-Order Effects of Neurons in CLIP

链接：https://iclr.cc/virtual/2025/poster/30282 abstract： We interpret the function of individual neurons in CLIP by automatically describing them using text. Analyzing the direct effects (i.e. the flow from a neuron through the residual stream to the output) or the indirect effects (overall contribution) fails to capture the neurons' function in CLIP. Therefore, we present the "second-order lens", analyzing the effect flowing from a neuron through the later attention heads, directly to the output. We find that these effects are highly selective: for each neuron, the effect is significant for <2% of the images. Moreover, each effect can be approximated by a single direction in the text-image space of CLIP. We describe neurons by decomposing these directions into sparse sets of text representations. The sets reveal polysemantic behavior - each neuron corresponds to multiple, often unrelated, concepts (e.g. ships and cars). Exploiting this neuron polysemy, we mass-produce "semantic" adversarial examples by generating

images with concepts spuriously correlated to the incorrect class. Additionally, we use the second-order effects for zero-shot segmentation, outperforming previous methods. Our results indicate that a automated interpretation of neurons can be used for model deception and for introducing new model capabilities

780. Interpreting and Editing Vision-Language Representations to Mitigate Hallucinations

链接：https://iclr.cc/virtual/2025/poster/30724 abstract：

781. Selective Attention Improves Transformer

链接：https://iclr.cc/virtual/2025/poster/27940 abstract： Unneeded elements in the attention’s context degrade performance. We introduce Selective Attention, a simple parameter-free change to the standard attention mechanism which reduces attention to unneeded elements. Selective attention consistently improves language modeling and downstream task performance in a variety of model sizes and context lengths. For example, transformers trained with the language modeling objective on C4 with selective attention perform language modeling equivalently to standard transformers with ~2X more heads and parameters in their attention modules. Selective attention also allows decreasing the size of the attention’s context buffer, leading to meaningful reductions in the memory and compute requirements during inference. For example, transformers trained on C4 with context sizes of 512, 1,024, and 2,048 need 16X, 25X, and 47X less memory for their attention module, respectively, when equipped with selective attention, as those without selective attention, with the same validation perplexity.

782. Locally Connected Echo State Networks for Time Series Forecasting

链接：https://iclr.cc/virtual/2025/poster/30041 abstract： Echo State Networks (ESNs) are a class of recurrent neural networks in which only a small readout regression layer is trained, while the weights of the recurrent network, termed the reservoir, are randomly assigned and remain fixed. Our work introduces the Locally Connected ESN (LCESN), a novel ESN variant with a locally connected reservoir, forced memory, and a weight adaptation strategy. LCESN significantly reduces the asymptotic time and space complexities compared to the conventional ESN, enabling substantially larger networks. LCESN also improves the memory properties of ESNs without affecting network stability. We evaluate LCESN's performance on the NARMA10 benchmark task and compare it to state-of-the-art models on nine real-world datasets. Despite the simplicity of our model and its one-shot training approach, LCESN achieves competitive results, even surpassing several state-of-the-art models. LCESN introduces a fresh approach to real-world time series forecasting and demonstrates that large, well-tuned random recurrent networks can rival complex gradient-trained feedforward models. We provide our GPU-based implementation of LCESN as an open-source library.

783. Knowledge Entropy Decay during Language Model Pretraining Hinders New Knowledge Acquisition

链接：https://iclr.cc/virtual/2025/poster/28937 abstract： In this work, we investigate how a model's tendency to broadly integrate its parametric knowledge evolves throughout pretraining, and how this behavior affects overall performance, particularly in terms of knowledge acquisition and forgetting. We introduce the concept of knowledge entropy, which quantifies the range of memory sources the model engages with; high knowledge entropy indicates that the model utilizes a wide range of memory sources, while low knowledge entropy suggests reliance on specific sources with greater certainty. Our analysis reveals a consistent decline in knowledge entropy as pretraining advances. We also find that the decline is closely associated with a reduction in the model's ability to acquire and retain knowledge, leading us to conclude that diminishing knowledge entropy (smaller number of active memory sources) impairs the model's knowledge acquisition and retention capabilities. We find further support for this by demonstrating that increasing the activity of inactive memory sources enhances the model's capacity for knowledge acquisition and retention.

784. Range, not Independence, Drives Modularity in Biologically Inspired Representations

链接：https://iclr.cc/virtual/2025/poster/30538 abstract： Why do biological and artificial neurons sometimes modularise, each encoding a single meaningful variable, and sometimes entangle their representation of many variables? In this work, we develop a theory of when biologically inspired networks---those that are nonnegative and energy efficient---modularise their representation of source variables (sources). We derive necessary and sufficient conditions on a sample of sources that determine whether the neurons in an optimal biologically-inspired linear autoencoder modularise. Our theory applies to any dataset, extending far beyond the case of statistical independence studied in previous work. Rather we show that sources modularise if their support is ``sufficiently spread''. From this theory, we extract and validate predictions in a variety of empirical studies on how data distribution affects modularisation in nonlinear feedforward and recurrent neural networks trained on supervised and unsupervised tasks. Furthermore, we apply these ideas to neuroscience data, showing that range independence can be used to understand the mixing or modularising of spatial and reward information in entorhinal recordings in seemingly conflicting experiments. Further, we use these results to suggest alternate origins of mixed-selectivity, beyond the predominant theory of flexible nonlinear classification. In sum, our theory prescribes precise conditions on when neural activities modularise,

providing tools for inducing and elucidating modular representations in brains and machines.

785. HexGen-2: Disaggregated Generative Inference of LLMs in Heterogeneous Environment

链接：https://iclr.cc/virtual/2025/poster/30492 abstract： Disaggregating the prefill and decoding phases represents an effective new paradigm for generative inference of large language models (LLM). This approach offers some significant system advantages, such as eliminating prefill-decoding interference and optimizing resource allocation. However, it is still an challenging open problem about how to deploy the disaggregated inference paradigm across a group of heterogeneous GPUs, which can be an economic alternative of the deployment over the homogeneous high performance GPUs.Towards this end, we introduce HexGen-2, a distributed system for high throughput and cost-efficient LLM serving on heterogeneous GPUs following the disaggragated paradigm. Built on top of HexGen, the core component of HexGen-2 is a sophisticated scheduling algorithm that formalizes the allocation of disaggregated LLM inference computations and communications over heterogeneous GPUs and network connections as a constraint optimization problem. We leverage the graph partitioning and max-flow algorithm to co-optimize resource allocation, parallel strategies for distinct inference phases, and the efficiency of inter-phase key-value (KV) cache communications. We conduct extensive experiments to evaluate HexGen-2, i.e., on OPT (30B) and Llama-2 (70B) models in various real-world settings, the results reveal that HexGen-2 delivers up to a 2.0$\times$ and on average a 1.3$\times$ improvement in serving throughput, reduces the average inference latency by 1.5$\times$ compared with state-of-the-art systems given the same price budget, and achieves comparable inference performance with a 30% lower price budget.

786. Stabilized Neural Prediction of Potential Outcomes in Continuous Time

链接：https://iclr.cc/virtual/2025/poster/29172 abstract： Patient trajectories from electronic health records are widely used to estimate conditional average potential outcomes (CAPOs) of treatments over time, which then allows to personalize care. Yet, existing neural methods for this purpose have a key limitation: while some adjust for time-varying confounding, these methods assume that the time series are recorded in discrete time. In other words, they are constrained to settings where measurements and treatments are conducted at fixed time steps, even though this is unrealistic in medical practice. In this work, we aim to estimate CAPOs in continuous time. The latter is of direct practical relevance because it allows for modeling patient trajectories where measurements and treatments take place at arbitrary, irregular timestamps. We thus propose a new method called stabilized continuous time inverse propensity network (SCIP-Net). For this, we further derive stabilized inverse propensity weights for robust estimation of the CAPOs. To the best of our knowledge, our SCIP-Net is the first neural method that performs proper adjustments for time-varying confounding in continuous time.

787. Model-agnostic meta-learners for estimating heterogeneous treatment effects over time

链接：https://iclr.cc/virtual/2025/poster/29699 abstract： Estimating heterogeneous treatment effects (HTEs) over time is crucial in many disciplines such as personalized medicine. Existing works for this task have mostly focused on model-based learners that adapt specific machine-learning models and adjustment mechanisms. In contrast, model-agnostic learners - so-called meta-learners - are largely unexplored. In our paper, we propose several meta-learners that are model-agnostic and thus can be used in combination with arbitrary machine learning models (e.g., transformers) to estimate HTEs over time. We then provide a comprehensive theoretical analysis that characterizes the different learners and that allows us to offer insights into when specific learners are preferable. Furthermore, we propose a novel IVW-DR-learner that (i) uses a doubly robust (DR) and orthogonal loss; and (ii) leverages inverse-variance weights (IVWs) that we derive to stabilize the DR-loss over time. Our IVWs downweight extreme trajectories due to products of inverse-propensities in the DR-loss, resulting in a lower estimation variance. Our IVW-DR-learner achieves superior performance in our experiments, particularly in regimes with low overlap and long time horizons.

788. Feature Responsiveness Scores: Model-Agnostic Explanations for Recourse

链接：https://iclr.cc/virtual/2025/poster/27815 abstract： Machine learning models routinely automate decisions in applications like lending and hiring. In such settings, consumer protection rules require companies that deploy models to explain predictions to decision subjects. These rules are motivated, in part, by the belief that explanations can promote recourse by revealing information that individuals can use to contest or improve their outcomes. In practice, many companies comply with these rules by providing individuals with a list of the most important features for their prediction, which they identify based on feature importance scores from feature attribution methods such as SHAP or LIME. In this work, we show how these practices can undermine consumers by highlighting features that would not lead to an improved outcome and by explaining predictions that cannot be changed. We propose to address these issues by highlighting features based on their responsiveness score—i.e., the probability that an individual can attain a target prediction by changing a specific feature. We develop efficient methods to compute responsiveness scores for any model and any dataset. We conduct an extensive empirical study on the responsiveness of explanations in lending. Our results show that standard practices in consumer finance can backfire by presenting consumers with reasons without recourse, and demonstrate how our approach improves consumer protection by highlighting responsive features and identifying fixed predictions.

789. Signature Kernel Conditional Independence Tests in Causal Discovery for Stochastic Processes

链接：https://iclr.cc/virtual/2025/poster/29848 abstract： Inferring the causal structure underlying stochastic dynamical systems from observational data holds great promise in domains ranging from science and health to finance. Such processes can often be accurately modeled via stochastic differential equations (SDEs), which naturally imply causal relationships via `which variables enter the differential of which other variables'. In this paper, we develop conditional independence (CI) constraints on coordinate processes over selected intervals that are Markov with respect to the acyclic dependence graph (allowing self-loops) induced by a general SDE model. We then provide a sound and complete causal discovery algorithm, capable of handling both fully and partially observed data, and uniquely recovering the underlying or induced ancestral graph by exploiting time directionality assuming a CI oracle. Finally, to make our algorithm practically usable, we also propose a flexible, consistent signature kernel-based CI test to infer these constraints from data. We extensively benchmark the CI test in isolation and as part of our causal discovery algorithms, outperforming existing approaches in SDE models and beyond.

790. Interpretable Causal Representation Learning for Biological Data in the Pathway Space

链接：https://iclr.cc/virtual/2025/poster/31089 abstract： Predicting the impact of genomic and drug perturbations in cellular function is crucial for understanding gene functions and drug effects, ultimately leading to improved therapies. To this end, Causal Representation Learning (CRL) constitutes one of the most promising approaches, as it aims to identify the latent factors that causally govern biological systems, thus facilitating the prediction of the effect of unseen perturbations. Yet, current CRL methods fail in reconciling their principled latent representations with known biological processes, leading to models that are not interpretable. To address this major issue, in this work we present SENA-discrepancy-VAE, a model based on the recently proposed CRL method discrepancy-VAE, that produces representations where each latent factor can be interpreted as the (linear) combination of the activity of a (learned) set of biological processes. To this extent, we present an encoder, SENA-δ, that efficiently compute and map biological processes' activity levels to the latent causal factors. We show that SENA-discrepancy-VAE achieves predictive performances on unseen combinations of interventions that are comparable with its original, non-interpretable counterpart, while inferring causal latent factors that are biologically meaningful.

791. Probabilistic Geometric Principal Component Analysis with application to neural data

链接：https://iclr.cc/virtual/2025/poster/28454 abstract： Dimensionality reduction is critical across various domains of science including neuroscience. Probabilistic Principal Component Analysis (PPCA) is a prominent dimensionality reduction method that provides a probabilistic approach unlike the deterministic approach of PCA and serves as a connection between PCA and Factor Analysis (FA). Despite their power, PPCA and its extensions are mainly based on linear models and can only describe the data in a Euclidean coordinate system around the mean of data. However, in many neuroscience applications, data may be distributed around a nonlinear geometry (i.e., manifold) rather than lying in the Euclidean space around the mean. We develop Probabilistic Geometric Principal Component Analysis (PGPCA) for such datasets as a new dimensionality reduction algorithm that can explicitly incorporate knowledge about a given nonlinear manifold that is first fitted from these data. Further, we show how in addition to the Euclidean coordinate system, a geometric coordinate system can be derived for the manifold to capture the deviations of data from the manifold and noise. We also derive a data-driven EM algorithm for learning the PGPCA model parameters. As such, PGPCA generalizes PPCA to better describe data distributions by incorporating a nonlinear manifold geometry. In simulations and brain data analyses, we show that PGPCA can effectively model the data distribution around various given manifolds and outperforms PPCA for such data. Moreover, PGPCA provides the capability to test whether the new geometric coordinate system better describes the data than the Euclidean one. Finally, PGPCA can perform dimensionality reduction and learn the data distribution both around and on the manifold. These capabilities make PGPCA valuable for enhancing the efficacy of dimensionality reduction for analysis of high-dimensional data that exhibit noise and are distributed around a nonlinear manifold, especially for neural data.

792. Data Scaling Laws in Imitation Learning for Robotic Manipulation

链接：https://iclr.cc/virtual/2025/poster/28305 abstract： Data scaling has revolutionized fields like natural language processing and computer vision, providing models with remarkable generalization capabilities. In this paper, we investigate whether similar data scaling laws exist in robotics, particularly in robotic manipulation, and whether appropriate data scaling can yield single-task robot policies that can be deployed zero-shot for any object within the same category in any environment. To this end, we conduct a comprehensive empirical study on data scaling in imitation learning. By collecting data across numerous environments and objects, we study how a policy’s generalization performance changes with the number of training environments, objects, and demonstrations. Throughout our research, we collect over 40,000 demonstrations and execute more than 15,000 real-world robot rollouts under a rigorous evaluation protocol. Our findings reveal several intriguing results: the generalization performance of the policy follows a roughly power-law relationship with the number of environments and objects. The diversity of environments and objects is far more important than the absolute number of demonstrations; once the number of demonstrations per environment or object reaches a certain threshold, additional demonstrations have minimal effect. Based on these insights, we propose an efficient data collection strategy. With four data collectors working for one afternoon, we collect

sufficient data to enable the policies for two tasks to achieve approximately 90\% success rates in novel environments with unseen objects.

793. TANGO: Co-Speech Gesture Video Reenactment with Hierarchical Audio Motion Embedding and Diffusion Interpolation

链接：https://iclr.cc/virtual/2025/poster/32094 abstract： We present TANGO, a framework for generating co-speech body-gesture videos. Given a few-minute, single-speaker reference video and target speech audio, TANGO produces high-fidelity videos with synchronized body gestures. TANGO builds on Gesture Video Reenactment (GVR), which splits and retrieves video clips using a directed graph structure - representing video frames as nodes and valid transitions as edges. We address two key limitations of GVR: audio-motion misalignment and visual artifacts in GAN-generated transition frames. In particular, i) we propose retrieving gestures using latent feature distance to improve cross-modal alignment. To ensure the latent features could effectively model the relationship between speech audio and gesture motion, we implement a hierarchical joint embedding space (AuMoClip); ii) we introduce the diffusion-based model to generate high-quality transition frames. Our diffusion model, Appearance Consistent Interpolation (ACInterp), is built upon AnimateAnyone and includes a reference motion module and homography background flow to preserve appearance consistency between generated and reference videos. By integrating these components into the graph-based retrieval framework, TANGO reliably produces realistic, audio-synchronized videos and outperforms all existing generative and retrieval methods. Our code, pretrained models, and datasets are publicly available at https://github.com/CyberAgentAILab/TANGO.

794. Can LLMs Separate Instructions From Data? And What Do We Even Mean By That?

链接：https://iclr.cc/virtual/2025/poster/30777 abstract： Large Language Models (LLMs) show impressive results in numerous practical applications, but they lack essential safety features that are common in other areas of computer science, particularly an explicit separation of instructions and data. This makes them vulnerable to manipulations such as indirect prompt injections and generally unsuitable for safety-critical tasks. Surprisingly, there is currently no established definition or benchmark to quantify this phenomenon. In this work, we close this gap by introducing a formal measure for instruction-data separation for single-turn language models and an empirical variant that is calculable from a model’s outputs. We also present a new dataset, SEP, that allows estimating the measure for real-world models. Our results on various LLMs show that the problem of instruction-data separation is real: all models fail to achieve high separation, and canonical mitigation techniques, such as prompt engineering and fine-tuning, either fail to substantially improve separation or reduce model utility.

795. DAWN: Dynamic Frame Avatar with Non-autoregressive Diffusion Framework for Talking head Video Generation

链接：https://iclr.cc/virtual/2025/poster/27895 abstract： Talking head generation intends to produce vivid and realistic talking head videos from a single portrait and speech audio clip. Although significant progress has been made in diffusion-based talking head generation, almost all methods rely on autoregressive strategies, which suffer from limited context utilization beyond the current generation step, error accumulation, and slower generation speed. To address these challenges, we present DAWN (\textbf{D}ynamic frame \textbf{A}vatar \textbf{W}ith \textbf{N}on-autoregressive diffusion), a framework that enables all-at-once generation of dynamic-length video sequences. Specifically, it consists of two main components: (1) audio-driven holistic facial dynamics generation in the latent motion space, and (2) audio-driven head pose and blink generation. Extensive experiments demonstrate that our method generates authentic and vivid videos with precise lip motions, and natural pose/blink movements. Additionally, with a high generation speed, DAWN possesses strong extrapolation capabilities, ensuring the stable production of high-quality long videos. These results highlight the considerable promise and potential impact of DAWN in the field of talking head video generation. Furthermore, we hope that DAWN sparks further exploration of non-autoregressive approaches in diffusion models. Our code will be publicly available at \url{https://github.com/Hanbo-Cheng/DAWN-pytorch}.

796. MCNC: Manifold-Constrained Reparameterization for Neural Compression

链接：https://iclr.cc/virtual/2025/poster/29420 abstract： The outstanding performance of large foundational models across diverse tasks,from computer vision to speech and natural language processing, has significantlyincreased their demand. However, storing and transmitting these models posessignificant challenges due to their massive size (e.g., 750GB for Llama 3.1 405B).Recent literature has focused on compressing the original weights or reducing thenumber of parameters required for fine-tuning these models. These compressionmethods generally constrain the parameter space, for example, through low-rankreparametrization (e.g., LoRA), pruning, or quantization (e.g., QLoRA) duringor after the model training. In this paper, we present a novel model compres-sion method, which we term Manifold-Constrained Neural Compression (MCNC).This method constrains the parameter space to low-dimensional pre-defined andfrozen nonlinear manifolds, which effectively cover this space. Given the preva-lence of good solutions in over-parameterized deep neural networks, we show thatby constraining the parameter space to our proposed manifold, we can identifyhigh-quality solutions while achieving unprecedented compression rates acrossa wide variety of tasks and architectures. Through extensive experiments incomputer vision and natural language processing tasks, we demonstrate that ourmethod significantly outperforms state-of-the-art baselines in terms of compres-sion,

accuracy, and/or model reconstruction time. Our code is publicly available athttps://github.com/mint-vu/MCNC.

797. Robust Conformal Prediction with a Single Binary Certificate

链接：https://iclr.cc/virtual/2025/poster/28494 abstract： Conformal prediction (CP) converts any model's output to prediction sets with a guarantee to cover the true label with (adjustable) high probability. Robust CP extends this guarantee to worst-case (adversarial) inputs. Existing baselines achieve robustness by bounding randomly smoothed conformity scores. In practice, they need expensive Monte-Carlo (MC) sampling (e.g. $\sim10^4$ samples per point) to maintain an acceptable set size. We propose a robust conformal prediction that produces smaller sets even with significantly lower MC samples (e.g. 150 for CIFAR10). Our approach binarizes samples with an adjustable (or automatically adjusted) threshold selected to preserve the coverage guarantee. Remarkably, we prove that robustness can be achieved by computing only one binary certificate, unlike previous methods that certify each calibration (or test) point. Thus, our method is faster and returns smaller robust sets. We also eliminate a previous limitation that requires a bounded score function.

798. Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning

链接：https://iclr.cc/virtual/2025/poster/30649 abstract： A promising approach for improving reasoning in large language models is to use process reward models (PRMs). PRMs provide feedback at each step of a multi-step reasoning trace, improving credit assignment over outcome reward models (ORMs) that only provide feedback at the final step. However, collecting dense, per-step human labels is not scalable, and training PRMs from automatically-labeled data has thus far led to limited gains. With the goal of using PRMs to improve a base policy via test-time search and reinforcement learning (RL), we ask: ``How should we design process rewards?'' Our key insight is that, to be effective, the process reward for a step should measure progress: a change in the likelihood of producing a correct response in the future, before and after taking the step, as measured under a prover policy distinct from the base policy. Such progress values can {distinguish} good and bad steps generated by the base policy, even though the base policy itself cannot. Theoretically, we show that even weaker provers can improve the base policy, as long as they distinguish steps without being too misaligned with the base policy. Our results show that process rewards defined as progress under such provers improve the efficiency of exploration during test-time search and online RL. We empirically validate our claims by training process advantage verifiers (PAVs) to measure progress under such provers and show that compared to ORM, they are >8% more accurate, and 1.5-5x more compute-efficient. Equipped with these insights, our PAVs enable one of the first results showing a 6x gain in sample efficiency for a policy trained using online RL with PRMs vs. ORMs.

799. RA-TTA: Retrieval-Augmented Test-Time Adaptation for Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/29434 abstract： Vision-language models (VLMs) are known to be susceptible to distribution shifts between pre-training data and test data, and test-time adaptation (TTA) methods for VLMs have been proposed to mitigate the detrimental impact of the distribution shifts. However, the existing methods solely rely on the internal knowledge encoded within the model parameters, which are constrained to pre-training data. To complement the limitation of the internal knowledge, we propose Retrieval-Augmented-TTA (RA-TTA) for adapting VLMs to test distribution using external knowledge obtained from a web-scale image database. By fully exploiting the bi-modality of VLMs, RA-TTA adaptively retrieves proper external images for each test image to refine VLMs' predictions using the retrieved external images, where fine-grained text descriptions are leveraged to extend the granularity of external knowledge. Extensive experiments on 17 datasets demonstrate that the proposed RA-TTA outperforms the state-of-the-art methods by 3.01-9.63\% on average.

800. Adaptive Rank Allocation: Speeding Up Modern Transformers with RaNA Adapters

链接：https://iclr.cc/virtual/2025/poster/27996 abstract： Large Language Models (LLMs) are computationally intensive, particularly during inference. Neuron-adaptive techniques, which selectively activate neurons in Multi-Layer Perceptron (MLP) layers, offer some speedups but suffer from limitations in modern Transformers. These include reliance on sparse activations, incompatibility with attention layers, and the use of costly neuron masking techniques. To address these issues, we propose the Adaptive Rank Allocation framework and introduce the Rank and Neuron Allocator (RaNA) adapter. RaNA adapters leverage rank adapters, which operate on linear layers by applying both low-rank matrix decompositions and adaptive masking to efficiently allocate compute without depending on activation sparsity. This enables RaNA to be generally applied to MLPs and linear components of attention modules, while eliminating the need for expensive maskers found in neuron-adaptive methods. Notably, when compared to neuron adapters, RaNA improves perplexity by up to 7 points and increases accuracy by up to 8 percentage-points when reducing FLOPs by $\sim$44\% in state-of-the-art Transformer architectures. These results position RaNA as a robust solution for improving inference efficiency in modern Transformer architectures.