1801. Private Mechanism Design via Quantile Estimation

链接：https://iclr.cc/virtual/2025/poster/30113 abstract： We investigate the problem of designing differentially private (DP), revenue-maximizing single item auction. Specifically, we consider broadly applicable settings in mechanism design where agents' valuation distributions are **independent**, **non-identical**, and can be either **bounded** or **unbounded**. Our goal is to design such auctions with **pure**, i.e., $(\epsilon,0)$ privacy in polynomial time. In this paper, we propose two computationally efficient auction learning framework that achieves **pure** privacy under bounded and unbounded distribution settings. These frameworks reduces the problem of privately releasing a revenue-maximizing auction to the private estimation of pre-specified quantiles. Our solutions increase the running time by polylog factors compared to the non-private version. As an application, we show how to extend our results to the multi-round online auction setting with non-myopic bidders. To our best knowledge, this paper is the first to efficiently deliver a Myerson auction with **pure** privacy and near-optimal revenue, and the first to provide such auctions for **unbounded** distributions.

1802. SAFREE: Training-Free and Adaptive Guard for Safe Text-to-Image And Video Generation

链接：https://iclr.cc/virtual/2025/poster/28752 abstract： Recent advances in diffusion models have significantly enhanced their ability to generate high-quality images and videos, but they have also increased the risk of producing unsafe content. Existing unlearning/editing-based methods for safe generation remove harmful concepts from models but face several challenges: (1) They cannot instantly remove harmful or undesirable concepts (e.g., artist styles) without additional training. (2) Their safe generation capabilities depend on collected training data. (3) They alter model weights, risking degradation in quality for content unrelated to the targeted toxic concepts. To address these challenges, we propose SAFREE, a novel, training-free approach for safe text-to-image and video generation, that does not alter the model's weights. Specifically, we detect a subspace corresponding to a set of toxic concepts in the text embedding space and steer prompt token embeddings away from this subspace, thereby filtering out harmful content while preserving intended semantics. To balance the trade-off between filtering toxicity and preserving safe concepts, SAFREE incorporates a novel self-validating filtering mechanism that dynamically adjusts the denoising steps when applying the filtered embeddings. Additionally, we incorporate adaptive re-attention mechanisms within the diffusion latent space to selectively diminish the influence of features related to toxic concepts at the pixel level. By integrating filtering across both textual embedding and visual latent spaces, SAFREE ensures coherent safety checking, preserving the fidelity, quality, and safety of the generated outputs. Empirically, SAFREE achieves state-of-the-art performance in suppressing unsafe content in T2I generation (reducing it by 22% across 5 datasets) compared to other training-free methods and effectively filters targeted concepts, e.g., specific artist styles, while maintaining high-quality output. It also shows competitive results against training-based methods. We further extend SAFREE to various T2I backbones and T2V tasks, showcasing its flexibility and generalization. As generative AI rapidly evolves, SAFREE provides a robust and adaptable safeguard for ensuring safe visual generation.

1803. Implicit Bias of Mirror Flow for Shallow Neural Networks in Univariate Regression

链接：https://iclr.cc/virtual/2025/poster/30180 abstract： We examine the implicit bias of mirror flow in least squares error regression with wide and shallow neural networks. For a broad class of potential functions, we show that mirror flow exhibits lazy training and has the same implicit bias as ordinary gradient flow when the network width tends to infinity. For univariate ReLU networks, we characterize this bias through a variational problem in function space. Our analysis includes prior results for ordinary gradient flow as a special case and lifts limitations which required either an intractable adjustment of the training data or networks with skip connections. We further introduce \emph{scaled potentials} and show that for these, mirror flow still exhibits lazy training but is not in the kernel regime. For univariate networks with absolute value activations, we show that mirror flow with scaled potentials induces a rich class of biases, which generally cannot be captured by an RKHS norm. A takeaway is that whereas the parameter initialization determines how strongly the curvature of the learned function is penalized at different locations of the input space, the scaled potential determines how the different magnitudes of the curvature are penalized.

1804. Multi-Label Test-Time Adaptation with Bound Entropy Minimization

链接：https://iclr.cc/virtual/2025/poster/30850 abstract： Mainstream test-time adaptation (TTA) techniques endeavor to mitigate distribution shifts via entropy minimization for multi-class classification, inherently increasing the probability of the most confident class. However, when encountering multi-label instances, the primary challenge stems from the varying number of labels per image, and prioritizing only the highest probability class inevitably undermines the adaptation of other positive labels. To address this issue, we investigate TTA within multi-label scenario (ML--TTA), developing Bound Entropy Minimization (BEM) objective to simultaneously increase the confidence of multiple top predicted labels. Specifically, to determine the number of labels for each augmented view, we retrieve a paired caption with yielded textual labels for that view. These labels are allocated to both the view and caption, called weak label set and strong label set with the same size k. Following this, the proposed BEM considers the highest top-k predicted labels from view and caption as a single entity, respectively, learning both view and caption prompts concurrently. By binding top-k predicted labels, BEM overcomes the limitation of vanilla entropy minimization, which exclusively optimizes the most confident class. Across the MSCOCO, VOC, and NUSWIDE multi-label datasets, our ML--TTA framework equipped with BEM exhibits superior performance compared to the latest SOTA methods, across various model architectures, prompt initialization, and varying label scenarios. The code is available at https://github.com/Jinx630/ML-

TTA.

1805. Subtask-Aware Visual Reward Learning from Segmented Demonstrations

链接：https://iclr.cc/virtual/2025/poster/28446 abstract： Reinforcement Learning (RL) agents have demonstrated their potential across various robotic tasks. However, they still heavily rely on human-engineered reward functions, requiring extensive trial-and-error and access to target behavior information, often unavailable in real-world settings. This paper introduces REDS: REward learning from Demonstration with Segmentations, a novel reward learning framework that leverages action-free videos with minimal supervision. Specifically, REDS employs video demonstrations segmented into subtasks from diverse sources and treats these segments as ground-truth rewards. We train a dense reward function conditioned on video segments and their corresponding subtasks to ensure alignment with ground-truth reward signals by minimizing the Equivalent-Policy Invariant Comparison distance. Additionally, we employ contrastive learning objectives to align video representations with subtasks, ensuring precise subtask inference during online interactions. Our experiments show that REDS significantly outperforms baseline methods on complex robotic manipulation tasks in Meta-World and more challenging real-world tasks, such as furniture assembly in FurnitureBench, with minimal human intervention. Moreover, REDS facilitates generalization to unseen tasks and robot embodiments, highlighting its potential for scalable deployment in diverse environments.

1806. Slot-Guided Adaptation of Pre-trained Diffusion Models for Object-Centric Learning and Compositional Generation

链接：https://iclr.cc/virtual/2025/poster/28572 abstract： We present SlotAdapt, an object-centric learning method that combines slot attention with pretrained diffusion models by introducing adapters for slot-based conditioning. Our method preserves the generative power of pretrained diffusion models, while avoiding their text-centric conditioning bias. We also incorporate an additional guidance loss into our architecture to align cross-attention from adapter layers with slot attention. This enhances the alignment of our model with the objects in the input image without using external supervision. Experimental results show that our method outperforms state-of-the-art techniques in object discovery and image generation tasks across multiple datasets, including those with real images. Furthermore, we demonstrate through experiments that our method performs remarkably well on complex real-world images for compositional generation, in contrast to other slot-based generative methods in the literature. The project page can be found at https://kaanakan.github.io/SlotAdapt/

1807. Trusted Multi-View Classification via Evolutionary Multi-View Fusion

链接：https://iclr.cc/virtual/2025/poster/29957 abstract： Multi-view classification based on the Dempster-Shafer theory is widely recognized for its reliability in safety-critical domains with multi-view data. However, the adoption of a late fusion strategy constrains information interaction among views, thereby leading to suboptimal utilization of multi-view data. A recent advancement addressing this limitation involves generating a pseudo view by concatenating individual views. Yet, the efficacy of this pseudo view may diminish when incorporating underperforming views like noisy views. Additionally, the integration of a pseudo view exacerbates the issue of imbalanced multi-view learning, as it contains a disproportionate amount of information compared to individual views. To address these issues, we propose the enhancing Trusted multi-view classification via Evolutionary multi-view Fusion (TEF) approach. TEF employs an evolutionary multi-view architecture search method to create a high-quality fusion architecture serving as the pseudo view, facilitating adaptive view and fusion operator selection. Furthermore, TEF enhances each view within the fusion architecture by concatenating the fusion architecture's decision output with its respective view. Our experimental results demonstrate the effectiveness of this straightforward yet powerful strategy in mitigating imbalanced multi-view learning issues, particularly on complex many-view datasets exceeding three views. Extensive evaluations across 13 multi-view datasets validate the superior performance of our proposed method compared to other trusted multi-view learning approaches. The code is available at https://github.com/fupinhan123/TEF.

1808. Federated Residual Low-Rank Adaption of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28955 abstract： Low-Rank Adaptation (LoRA) presents an effective solution for federated fine-tuning of Large Language Models (LLMs), as it substantially reduces communication overhead. However, a straightforward combination of FedAvg and LoRA results in suboptimal performance, especially under data heterogeneity. We noted this stems from both intrinsic (i.e., constrained parameter space) and extrinsic (i.e., client drift) limitations, which hinder it effectively learn global knowledge. In this work, we proposed a novel Federated Residual Low-Rank Adaption method, namely FRLoRA, to tackle above two limitations. It directly sums the weight of the global model parameters with a residual low-rank matrix product (\ie, weight change) during the global update step, and synchronizes this update for all local models. By this, FRLoRA performs global updates in a higher-rank parameter space, enabling a better representation of complex knowledge structure. Furthermore, FRLoRA reinitializes the local low-rank matrices with the principal singular values and vectors of the pre-trained weights in each round, to calibrate their inconsistent convergence, thereby mitigating client drift. Our extensive experiments demonstrate that FRLoRA consistently outperforms various state-of-the-art FL methods across nine different benchmarks in natural language understanding and generation under different FL scenarios.

1809. Flavors of Margin: Implicit Bias of Steepest Descent in Homogeneous

Neural Networks

链接：https://iclr.cc/virtual/2025/poster/30582 abstract： We study the implicit bias of the family of steepest descent algorithms with infinitesimal learning rate, including gradient descent, sign gradient descent and coordinate descent, in deep homogeneous neural networks. We prove that an algorithm-dependent geometric margin increases during training and characterize the late-stage bias of the algorithms. In particular, we define a generalized notion of stationarity for optimization problems and show that the algorithms progressively reduce a (generalized) Bregman divergence, which quantifies proximity to such stationary points of a margin-maximization problem. We then experimentally zoom into the trajectories of neural networks optimized with various steepest descent algorithms, highlighting connections to the implicit bias of Adam.

1810. On the Importance of Language-driven Representation Learning for Heterogeneous Federated Learning

链接：https://iclr.cc/virtual/2025/poster/30797 abstract： Non-Independent and Identically Distributed (Non-IID) training data significantly challenge federated learning (FL), impairing the performance of the global model in distributed frameworks. Inspired by the superior performance and generalizability of language-driven representation learning in centralized settings, we explore its potential to enhance FL for handling non-IID data. In specific, this paper introduces FedGLCL, a novel language-driven FL framework for image-text learning that uniquely integrates global language and local image features through contrastive learning, offering a new approach to tackle non-IID data in FL. FedGLCL redefines FL by avoiding separate local training models for each client. Instead, it uses contrastive learning to harmonize local image features with global textual data, enabling uniform feature learning across different local models. The utilization of a pre-trained text encoder in FedGLCL serves a dual purpose: it not only reduces the variance in local feature representations within FL by providing a stable and rich language context but also aids in mitigating overfitting, particularly to majority classes, by leveraging broad linguistic knowledge. Extensive experiments show that FedGLCL significantly outperforms state-of-the-art FL algorithms across different non-IID scenarios.

1811. Pareto Prompt Optimization

链接：https://iclr.cc/virtual/2025/poster/30234 abstract： Natural language prompt optimization, or prompt engineering, has emerged as a powerful technique to unlock the potential of Large Language Models (LLMs) for various tasks. While existing methods primarily focus on maximizing a single task-specific performance metric for LLM outputs, real-world applications often require considering trade-offs between multiple objectives. In this work, we address this limitation by proposing an effective technique for multi-objective prompt optimization for LLMs. Specifically, we propose ParetoPrompt, a reinforcement learning~ (RL) method that leverages dominance relationships between prompts to derive a policy model for prompts optimization using preference-based loss functions. By leveraging multi-objective dominance relationships, ParetoPrompt enables efficient exploration of the entire Pareto front without the need for a predefined scalarization of multiple objectives. Our experimental results show that ParetoPrompt consistently outperforms existing algorithms that use specific objective values. ParetoPrompt also yields robust performances when the objective metrics differ between training and testing.

1812. Revisiting Random Walks for Learning on Graphs

链接：https://iclr.cc/virtual/2025/poster/29618 abstract： We revisit a simple model class for machine learning on graphs, where a random walk on a graph produces a machine-readable record, and this record is processed by a deep neural network to directly make vertex-level or graph-level predictions. We call these stochastic machines random walk neural networks (RWNNs), and through principled analysis, show that we can design them to be isomorphism invariant while capable of universal approximation of graph functions in probability. A useful finding is that almost any kind of record of random walks guarantees probabilistic invariance as long as the vertices are anonymized. This enables us, for example, to record random walks in plain text and adopt a language model to read these text records to solve graph tasks. We further establish a parallelism to message passing neural networks using tools from Markov chain theory, and show that over-smoothing in message passing is alleviated by construction in RWNNs, while over-squashing manifests as probabilistic under-reaching. We empirically demonstrate RWNNs on a range of problems, verifying our theoretical analysis and demonstrating the use of language models for separating strongly regular graphs where 3-WL test fails, and transductive classification on arXiv citation network. Code is available at https://github.com/jw9730/random-walk.

1813. Algorithmic Stability Based Generalization Bounds for Adversarial Training

链接：https://iclr.cc/virtual/2025/poster/31151 abstract： In this paper, we present a novel stability analysis of adversarial training and prove generalization upper bounds in terms of an expansiveness property of adversarial perturbations used during training and used for evaluation. These expansiveness parameters appear not only govern the vanishing rate of the generalization error but also govern its scaling constant. Our proof techniques do not rely on artificial assumptions of the adversarial loss, as are typically used in previous works. Our bound attributes the robust overfitting in PGD-based adversarial training to the sign function used in the PGD attack, resulting in a bad expansiveness parameter. The peculiar choice of sign function in the PGD attack appears to impact adversarial training both in terms of (inner) optimization and in terms of generalization, as shown in this work. This aspect has been largely overlooked to date. Going beyond the sign-function based

PGD attacks, we further show that poor expansiveness properties exist in a wide family of PGD-like iterative attack algorithms, which may highlight an intrinsic difficulty in adversarial training.

1814. A Generalist Hanabi Agent

链接：https://iclr.cc/virtual/2025/poster/28311 abstract： Traditional multi-agent reinforcement learning (MARL) systems can develop cooperative strategies through repeated interactions. However, these systems are unable to perform well on any other setting than the one they have been trained on, and struggle to successfully cooperate with unfamiliar collaborators. This is particularly visible in the Hanabi benchmark, a popular 2-to-5 player cooperative card-game which requires complex reasoning and precise assistance to other agents. Current MARL agents for Hanabi can only learn one specific game-setting (e.g., 2-player games), and play with the same algorithmic agents. This is in stark contrast to humans, who can quickly adjust their strategies to work with unfamiliar partners or situations. In this paper, we introduce Recurrent Replay Relevance Distributed DQN (R3D2), a generalist agent for Hanabi, designed to overcome these limitations. We reformulate the task using text, as language has been shown to improve transfer. We then propose a distributed MARL algorithm that copes with the resulting dynamic observation- and action-space. In doing so, our agent is the first that can play all game settings concurrently, and extend strategies learned from one setting to other ones. As a consequence, our agent also demonstrates the ability to collaborate with different algorithmic agents ---agents that are themselves unable to do so.

1815. WebRL: Training LLM Web Agents via Self-Evolving Online Curriculum Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28354 abstract： Large language models (LLMs) have shown remarkable potential as autonomous agents, particularly in web-based tasks. However, existing LLM web agents face significant limitations: high-performing agents rely on expensive proprietary LLM APIs, while open LLMs lack the necessary decision-making capabilities. This paper introduces WebRL, a novel self-evolving online curriculum reinforcement learning framework designed to train high-performance web agents using open LLMs. Our approach addresses key challenges in this domain, including the scarcity of training tasks, sparse feedback signals, and policy distribution drift in online learning. WebRL incorporates a self-evolving curriculum that generates new tasks from unsuccessful attempts, a robust outcome-supervised reward model (ORM), and adaptive reinforcement learning strategies to ensure consistent improvement. We apply WebRL to transform Llama-3.1 models into proficient web agents, achieving remarkable results on the WebArena-Lite benchmark. Our Llama-3.1-8B agent improves from an initial 4.8\% success rate to 42.4\%, while the Llama-3.1-70B agent achieves a 47.3\% success rate across five diverse websites. These results surpass the performance of GPT-4-Turbo (17.6\%) by over 160\% relatively and significantly outperform previous state-of-the-art web agents trained on open LLMs (AutoWebGLM, 18.2\%). Our findings demonstrate WebRL's effectiveness in bridging the gap between open and proprietary LLM-based web agents, paving the way for more accessible and powerful autonomous web interaction systems.

1816. Following the Human Thread in Social Navigation

链接：https://iclr.cc/virtual/2025/poster/29952 abstract： The success of collaboration between humans and robots in shared environments relies on the robot's real-time adaptation to human motion. Specifically, in Social Navigation, the agent should be close enough to assist but ready to back up to let the human move freely, avoiding collisions. Human trajectories emerge as crucial cues in Social Navigation, but they are partially observable from the robot's egocentric view and computationally complex to process.We present the first Social Dynamics Adaptation model (SDA) based on the robot's state-action history to infer the social dynamics. We propose a two-stage Reinforcement Learning framework: the first learns to encode the human trajectories into social dynamics and learns a motion policy conditioned on this encoded information, the current status, and the previous action. Here, the trajectories are fully visible, i.e., assumed as privileged information. In the second stage, the trained policy operates without direct access to trajectories. Instead, the model infers the social dynamics solely from the history of previous actions and statuses in real-time.Tested on the novel Habitat 3.0 platform, SDA sets a novel state-of-the-art (SotA) performance in finding and following humans. The code can be found at https://github.com/L-Scofano/SDA.

1817. Population Transformer: Learning Population-level Representations of Neural Activity

链接：https://iclr.cc/virtual/2025/poster/30337 abstract： We present a self-supervised framework that learns population-level codes for arbitrary ensembles of neural recordings at scale. We address key challenges in scaling models with neural time-series data, namely, sparse and variable electrode distribution across subjects and datasets. The Population Transformer (PopT) stacks on top of pretrained temporal embeddings and enhances downstream decoding by enabling learned aggregation of multiple spatially-sparse data channels. The pretrained PopT lowers the amount of data required for downstream decoding experiments, while increasing accuracy, even on held-out subjects and tasks. Compared to end-to-end methods, this approach is computationally lightweight, while achieving similar or better decoding performance. We further show how our framework is generalizable to multiple time-series embeddings and neural data modalities. Beyond decoding, we interpret the pretrained and fine-tuned PopT models to show how they can be used to extract neuroscience insights from large amounts of data. We release our code as well as a pretrained PopT to enable off-the-shelf improvements in multi-channel intracranial data decoding and interpretability. Code is available at https://github.com/czlwang/PopulationTransformer.

1818. Strategist: Self-improvement of LLM Decision Making via Bi-Level Tree Search

链接：https://iclr.cc/virtual/2025/poster/28807 abstract： Traditional reinforcement learning and planning require a lot of data and training to develop effective strategies. On the other hand, large language models (LLMs) can generalize well and perform tasks without prior training but struggle with complex planning and decision-making. We introduce STRATEGIST, a new approach that combines the strengths of both methods. It uses LLMs to generate and update high-level strategies in text form, while a Monte Carlo Tree Search (MCTS) algorithm refines and executes them. STRATEGIST is a general framework that optimizes strategies through self-play simulations without requiring any training data. We test STRATEGIST in competitive, multi-turn games with partial information, such as Game of Pure Strategy (GOPS) and The Resistance: Avalon, a multi-agent hidden-identity discussion game. Our results show that STRATEGIST-based agents outperform traditional reinforcement learning models, other LLM-based methods, and existing LLM agents while achieving performance levels comparable to human players.

1819. CoTFormer: A Chain of Thought Driven Architecture with Budget-Adaptive Computation Cost at Inference

链接：https://iclr.cc/virtual/2025/poster/30808 abstract： Scaling language models to larger and deeper sizes has led to significant boosts in performance. Even though the size of these models limits their application in compute-constrained environments, the race to continually develop ever larger and deeper foundational models is underway. At the same time---regardless of the model size---task-specific techniques continue to play a pivotal role in achieving optimal downstream performance. One of these techniques, called Chain-of-Thought (CoT), is particularly interesting since, as we point out in this work, it resembles employing a deeper transformer through re-applying the model multiple times. However, a key subtlety in computing the attention of past tokens differentiates CoT from simply applying the model several times. Based on this insight, we propose CoTFormer, a novel architecture which closely mimics CoT at the token level, allowing us to obtain significantly improved accuracies close to much larger models. While applying CoT introduces additional computation costs, we compensate for it by leveraging CoTFormer's special compatibility with token-wise variable depth. Through a compute adaptive model---which automatically allocates the compute to tokens that need it most---we show that it is possible to reduce the computation cost significantly without any reduction in accuracy, and with further compute cost reductions possible while maintaining a competitive accuracy.

1820. LancBiO: Dynamic Lanczos-aided Bilevel Optimization via Krylov Subspace

链接：https://iclr.cc/virtual/2025/poster/27854 abstract： Bilevel optimization, with broad applications in machine learning, has an intricate hierarchical structure. Gradient-based methods have emerged as a common approach to large-scale bilevel problems. However, the computation of the hyper-gradient, which involves a Hessian inverse vector product, confines the efficiency and is regarded as a bottleneck. To circumvent the inverse, we construct a sequence of low-dimensional approximate Krylov subspaces with the aid of the Lanczos process. As a result, the constructed subspace is able to dynamically and incrementally approximate the Hessian inverse vector product with less effort and thus leads to a favorable estimate of the hyper-gradient. Moreover, we propose a provable subspace-based framework for bilevel problems where one central step is to solve a small-size tridiagonal linear system. To the best of our knowledge, this is the first time that subspace techniques are incorporated into bilevel optimization. This successful trial not only enjoys $\mathcal{O}(\epsilon^{-1})$ convergence rate but also demonstrates efficiency in a synthetic problem and two deep learning tasks.

1821. ReAttention: Training-Free Infinite Context with Finite Attention Scope

链接：https://iclr.cc/virtual/2025/poster/30071 abstract： The long-context capability of the Large Language Models (LLM) has made significant breakthroughs, but \textit{the maximum supported context length in length extrapolation} remains a critical bottleneck limiting their practical applications. The constraint of context length in LLMs arises from the self-attention mechanism, which cannot effectively and efficiently capture the semantic relationships within infinitely long contexts via the limited pre-trained positional information and attention scope. In this work, we propose \textbf{ReAttention}, a training-free approach enabling LLM based on the self-attention mechanism to support an infinite context with a finite attention scope under sufficient memory resources. ReAttention performs the position-agnostic top-k attention before the ordinary position-aware self-attention, freeing LLMs from the length extrapolation issue. We validate the performance of ReAttention on the LongBench, L-Eval, and InfiniteBench and demonstrate that it is on par with traditional methods. Furthermore, we also apply ReAttention on mainstream LLMs, including LLaMA3.1-8B and Mistral-v0.3-7B, enabling them to support context lengths of at least 1M and even expanding the context length of LLaMA3.2-3B-chat by 128$\times$ to 4M without any further training in Needle-In-A-Haystack tests. We also improve the efficiency of ReAttention with Triton and achieve an efficient extrapolation without additional overhead. The code is available at \url{https://github.com/OpenMOSS/ReAttention}.

1822. PCNN: Probable-Class Nearest-Neighbor Explanations Improve Fine-Grained Image Classification Accuracy for AIs and Humans

链接：https://iclr.cc/virtual/2025/poster/31474 abstract： Nearest neighbors (NN) are traditionally used to compute final decisions, e.g., in Support Vector Machines or k-NN classifiers, and to provide users with explanations for the model's decision. In this paper, we show a novel utility of nearest neighbors: To improve predictions of a frozen, pretrained image classifier C. We leverage an image comparator S that (1) compares the input image with NN images from the top-K most probable classes given by C; and (2) uses scores from S to weight the confidence scores of C to refine predictions. Our method consistently improves fine-grained image classification accuracy on CUB-200, Cars-196, and Dogs-120. Also, a human study finds that showing users our probable-class nearest neighbors (PCNN) reduces over-reliance on AI, thus improving their decision accuracy over prior work which only shows only the most-probable (top-1) class examples.

1823. Bringing NeRFs to the Latent Space: Inverse Graphics Autoencoder

链接：https://iclr.cc/virtual/2025/poster/29995 abstract： While pre-trained image autoencoders are increasingly utilized in computer vision, the application of inverse graphics in 2D latent spaces has been under-explored. Yet, besides reducing the training and rendering complexity, applying inverse graphics in the latent space enables a valuable interoperability with other latent-based 2D methods. The major challenge is that inverse graphics cannot be directly applied to such image latent spaces because they lack an underlying 3D geometry. In this paper, we propose an Inverse Graphics Autoencoder (IG-AE) that specifically addresses this issue. To this end, we regularize an image autoencoder with 3D-geometry by aligning its latent space with jointly trained latent 3D scenes. We utilize the trained IG-AE to bring NeRFs to the latent space with a latent NeRF training pipeline, which we implement in an open-source extension of the Nerfstudio framework, thereby unlocking latent scene learning for its supported methods. We experimentally confirm that Latent NeRFs trained with IG-AE present an improved quality compared to a standard autoencoder, all while exhibiting training and rendering accelerations with respect to NeRFs trained in the image space. Our project page can be found at https://ig-ae.github.io .

1824. Matrix Product Sketching via Coordinated Sampling

链接：https://iclr.cc/virtual/2025/poster/28936 abstract： We revisit the well-studied problem of approximating a matrix product, $\bv{A}^T\bv{B}$, based on small space sketches $\mathcal{S}(\bv{A})$ and $\mathcal{S}(\bv{B})$ of $\bv{A} \in \R^{n \times d}$ and $\bv{B}\in \R^{n \times m}$. We are interested in the setting where the sketches must be computed independently of each other, except for the use of a shared random seed. We prove that, when $\bv{A}$ and $\bv{B}$ are sparse, methods based on \emph{coordinated random sampling} can outperform classical linear sketching approaches, like Johnson-Lindenstrauss Projection or CountSketch. For example, to obtain Frobenius norm error $\epsilon\|\bv{A}\|_F\|\bv{B}\|_F$, coordinated sampling requires sketches of size $O(s/\epsilon^2)$ when $\bv{A}$ and $\bv{B}$ have at most $s \leq d,m$ non-zeros per row. In contrast, linear sketching leads to sketches of size $O(d/\epsilon^2)$ and $O(m/\epsilon^2)$ for $\bv{A}$ and $\bv{B}$. We empirically evaluate our approach on two applications: 1) distributed linear regression in databases, a problem motivated by tasks like dataset discovery and augmentation, and 2) approximating attention matrices in transformer-based language models. In both cases, our sampling algorithms yield an order of magnitude improvement over linear sketching.

1825. Uncovering Latent Memories in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30058 abstract： Frontier AI systems are making transformative impacts across society, but such benefits are not without costs: models trained on web-scale datasets containing personal and private data raise profound concerns about data privacy and security. Language models are trained on extensive corpora including potentially sensitive or proprietary information, and the risk of data leakage, where the model response reveals pieces of such information, remains inadequately understood. Prior work has investigated that sequence complexity and the number of repetitions are the primary drivers of memorization. In this work, we examine the most vulnerable class of data: highly complex sequences that are presented only once during training. These sequences often contain the most sensitive information and pose considerable risk if memorized. By analyzing the progression of memorization for these sequences throughout training, we uncover a striking observation: many memorized sequences persist in the model's memory, exhibiting resistance to catastrophic forgetting even after just one encounter. Surprisingly, these sequences may not appear memorized immediately after their first exposure but can later be “uncovered” during training, even in the absence of subsequent exposures - a phenomenon we call "latent memorization." Latent memorization presents a serious challenge for data privacy, as sequences that seem hidden at the final checkpoint of a model may still be easily recoverable. We demonstrate how these hidden sequences can be revealed through random weight perturbations, and we introduce a diagnostic test based on cross-entropy loss to accurately identify latent memorized sequences.

1826. PIORF: Physics-Informed Ollivier-Ricci Flow for Long–Range Interactions in Mesh Graph Neural Networks

链接：https://iclr.cc/virtual/2025/poster/28238 abstract： Recently, data-driven simulators based on graph neural networks have gained attention in modeling physical systems on unstructured meshes. However, they struggle with long-range dependencies in fluid flows, particularly in refined mesh regions. This challenge, known as the 'over-squashing' problem, hinders information propagation. While existing graph rewiring methods address this issue to some extent, they only consider graph topology, overlooking the underlying physical phenomena. We propose Physics-Informed Ollivier--Ricci Flow (PIORF), a novel rewiring method that combines physical correlations with graph topology. PIORF uses Ollivier--Ricci curvature (ORC) to identify bottleneck regions and connects these areas with nodes in high-velocity gradient nodes, enabling long-range interactions and mitigating over-squashing. Our approach is computationally efficient in rewiring edges and can scale to larger simulations.

Experimental results on 3 fluid dynamics benchmark datasets show that PIORF consistently outperforms baseline models and existing rewiring methods, achieving up to 26.2\% improvement.

1827. Learning Spatiotemporal Dynamical Systems from Point Process Observations

链接：https://iclr.cc/virtual/2025/poster/31095 abstract： Spatiotemporal dynamics models are fundamental for various domains, from heat propagation in materials to oceanic and atmospheric flows. However, currently available neural network-based spatiotemporal modeling approaches fall short when faced with data that is collected randomly over time and space, as is often the case with sensor networks in real-world applications like crowdsourced earthquake detection or pollution monitoring. In response, we developed a new method that can effectively learn spatiotemporal dynamics from such point process observations. Our model integrates techniques from neural differential equations, neural point processes, implicit neural representations and amortized variational inference to model both the dynamics of the system and the probabilistic locations and timings of observations. It outperforms existing methods on challenging spatiotemporal datasets by offering substantial improvements in predictive accuracy and computational efficiency, making it a useful tool for modeling and understanding complex dynamical systems observed under realistic, unconstrained conditions.

1828. NeRAF: 3D Scene Infused Neural Radiance and Acoustic Fields

链接：https://iclr.cc/virtual/2025/poster/28397 abstract： Sound plays a major role in human perception. Along with vision, it provides essential information for understanding our surroundings. Despite advances in neural implicit representations, learning acoustics that align with visual scenes remains a challenge. We propose NeRAF, a method that jointly learns acoustic and radiance fields. NeRAF synthesizes both novel views and spatialized room impulse responses (RIR) at new positions by conditioning the acoustic field on 3D scene geometric and appearance priors from the radiance field. The generated RIR can be applied to auralize any audio signal. Each modality can be rendered independently and at spatially distinct positions, offering greater versatility. We demonstrate that NeRAF generates high-quality audio on SoundSpaces and RAF datasets, achieving significant performance improvements over prior methods while being more data-efficient. Additionally, NeRAF enhances novel view synthesis of complex scenes trained with sparse data through cross-modal learning.NeRAF is designed as a Nerfstudio module, providing convenient access to realistic audio-visual generation. Project page: https://amandinebtto.github.io/NeRAF

1829. TimeInf: Time Series Data Contribution via Influence Functions

链接：https://iclr.cc/virtual/2025/poster/29389 abstract： Evaluating the contribution of individual data points to a model's prediction is critical for interpreting model predictions and improving model performance. Existing data contribution methods have been applied to various data types, including tabular data, images, and text; however, their primary focus has been on i.i.d. settings. Despite the pressing need for principled approaches tailored to time series datasets, the problem of estimating data contribution in such settings remains under-explored, possibly due to challenges associated with handling inherent temporal dependencies. This paper introduces TimeInf, a model-agnostic data contribution estimation method for time-series datasets. By leveraging influence scores, TimeInf attributes model predictions to individual time points while preserving temporal structures between the time points. Our empirical results show that TimeInf effectively detects time series anomalies and outperforms existing data attribution techniques as well as state-of-the-art anomaly detection methods. Moreover, TimeInf offers interpretable attributions of data values, allowing us to distinguish diverse anomalous patterns through visualizations. We also showcase a potential application of TimeInf in identifying mislabeled anomalies in the ground truth annotations.

1830. Trivialized Momentum Facilitates Diffusion Generative Modeling on Lie Groups

链接：https://iclr.cc/virtual/2025/poster/30456 abstract： The generative modeling of data on manifolds is an important task, for which diffusion models in flat spaces typically need nontrivial adaptations. This article demonstrates how a technique called `trivialization' can transfer the effectiveness of diffusion models in Euclidean spaces to Lie groups. In particular, an auxiliary momentum variable was algorithmically introduced to help transport the position variable between data distribution and a fixed, easy-to-sample distribution. Normally, this would incur further difficulty for manifold data because momentum lives in a space that changes with the position. However, our trivialization technique creates a new momentum variable that stays in a simple fixed vector space. This design, together with a manifold preserving integrator, simplifies implementation and avoids inaccuracies created by approximations such as projections to tangent space and manifold, which were typically used in prior work, hence facilitating generation with high-fidelity and efficiency. The resulting method achieves state-of-the-art performance on protein and RNA torsion angle generation and sophisticated torus datasets. We also, arguably for the first time, tackle the generation of data on high-dimensional Special Orthogonal and Unitary groups, the latter essential for quantum problems. Code is available at https://github.com/yuchen-zhu-zyc/TDM.

1831. Multi-Field Adaptive Retrieval

链接：https://iclr.cc/virtual/2025/poster/31079 abstract： Document retrieval for tasks such as search and retrieval-augmented generation typically involves datasets that are unstructured: free-form text without explicit internal structure in each document.

However, documents can have some structure, containing fields such as an article title, a message body, or an HTML header. To address this gap, we introduce Multi-Field Adaptive Retrieval (mFAR), a flexible framework that accommodates any number and any type of document indices on semi-structured data. Our framework consists of two main steps: (1) the decomposition of an existing document into fields, each indexed independently through dense and lexical methods, and (2) learning a model which adaptively predicts the importance of a field by conditioning on the document query, allowing on-the-fly weighting of the most likely field(s). We find that our approach allows for the optimized use of dense versus lexical representations across field types, significantly improves in document ranking over a number of existing retrievers, and achieves state-of-the-art performance for multi-field structured data.

1832. Diffusion Generative Modeling for Spatially Resolved Gene Expression Inference from Histology Images

链接：https://iclr.cc/virtual/2025/poster/30320 abstract：

1833. Controllable Safety Alignment: Inference-Time Adaptation to Diverse Safety Requirements

链接：https://iclr.cc/virtual/2025/poster/30405 abstract： The current paradigm for safety alignment of large language models (LLMs) follows a one-size-fits-all approach: the model refuses to interact with any content deemed unsafe by the model provider. This approach lacks flexibility in the face of varying social norms across cultures and regions. In addition, users may have diverse safety needs, making a model with static safety standards too restrictive to be useful, as well as too costly to be re-aligned.We propose Controllable Safety Alignment (CoSA), a framework designed to adapt models to diverse safety requirements without re-training. Instead of aligning a fixed model, we align models to follow safety configs—free-form natural language descriptions of the desired safety behaviors—that are provided as part of the system prompt. To adjust model safety behavior, authorized users only need to modify such safety configs at inference time. To enable that, we propose CoSAlign, a data-centric method for aligning LLMs to easily adapt to diverse safety configs. Furthermore, we devise a novel controllability evaluation protocol that considers both helpfulness and configured safety, summarizing them into CoSA-Score, and construct CoSApien, a human-authored benchmark that consists of real-world LLM use cases with diverse safety requirements and corresponding evaluation prompts. We show that CoSAlign leads to substantial gains of controllability over strong baselines including in-context alignment. Our framework encourages better representation and adaptation to pluralistic human values in LLMs, and thereby increasing their practicality.

1834. Learning General-purpose Biomedical Volume Representations using Randomized Synthesis

链接：https://iclr.cc/virtual/2025/poster/27787 abstract： Current volumetric biomedical foundation models struggle to generalize as public 3D datasets are small and do not cover the broad diversity of medical procedures, conditions, anatomical regions, and imaging protocols. We address this by creating a representation learning method that instead anticipates strong domain shifts at training time itself. We first propose a data engine that synthesizes highly variable training samples that would enable generalization to new biomedical contexts. To then train a single 3D network for any voxel-level task, we develop a contrastive learning method that pretrains the network to be stable against nuisance imaging variation simulated by the data engine, a key inductive bias for generalization. This network's features can be used as robust representations of input images for downstream tasks and its weights provide a strong, dataset-agnostic initialization for finetuning on new datasets. As a result, we set new standards across both multimodality registration and few-shot segmentation, a first for any 3D biomedical vision model, all without (pre-)training on any existing dataset of real images.

1835. Context Steering: Controllable Personalization at Inference Time

链接：https://iclr.cc/virtual/2025/poster/27781 abstract： To deliver high-quality, personalized responses, large language models (LLMs) must effectively incorporate context — personal, demographic, and cultural information specific to an end-user. For example, asking the model to explain Newton's second law with the context "I am a toddler" should produce a response different from when the context is "I am a physics professor". However, leveraging the context in practice is a nuanced and challenging task, and is often dependent on the specific situation or user base. The model must strike a balance between providing specific, personalized responses and maintaining general applicability. Current solutions, such as prompt-engineering and fine-tuning, require collection of contextually appropriate responses as examples, making them time-consuming and less flexible to use across different contexts. In this work, we introduce Context Steering (CoS) —a simple, training-free decoding approach that amplifies the influence of the context in next token predictions. CoS computes contextual influence by comparing the output probabilities from two LLM forward passes: one that includes the context and one that does not. By linearly scaling the contextual influence, CoS allows practitioners to flexibly control the degree of personalization for different use cases. We show that CoS can be applied to autoregressive LLMs, and demonstrates strong performance in personalized recommendations. Additionally, we show that CoS can function as a Bayesian Generative model to infer and quantify correlations between open-ended texts, broadening its potential applications.

1836. Generative Adapter: Contextualizing Language Models in Parameters

with A Single Forward Pass

链接：https://iclr.cc/virtual/2025/poster/29103 abstract： Large language models (LLMs) acquire substantial knowledge during pretraining but often need adaptation to new contexts, tasks, or domains, typically achieved through fine-tuning or prompting. However, fine-tuning incurs significant training costs, while prompting increases inference overhead. Inspired by fast weight memory, we introduce GenerativeAdapter, an effective and efficient adaptation method that encode test-time context into language model parameters with a single forward pass.GenerativeAdapter augments a frozen pretrained LM with a lightweight adapter generator, trained via self-supervised learning, to produce parameter-efficient adapters.Notably, our generator is general-purpose, i.e., one generator can adapt the corresponding base model for all langauge processing scenarios.We apply GenerativeAdapter to two pretrained LMs (Mistral-7B-Instruct and Llama2-7B-Chat) and evaluate the adapted models across knowledge acquisition from documents, learning from demonstrations, and personalization for users.In StreamingQA, our approach is effective in injecting knowledge into the LM's parameters, achieving a 63.5\% improvement in F1 score over the model with supervised fine-tuning (from 19.5 to 31.5) for contexts as long as 32K tokens.In the MetaICL in-context learning evaluation, our method achieves an average accuracy of 44.9 across 26 tasks, outperforming the base model. On MSC, our method proves to be highly competitive in memorizing user information from conversations with a 4x reduction in computation and memory costs compared to prompting with full conversation history.Overall, GenerativeAdapter provides a viable solution for adapting large LMs to evolving information and providing tailored user experience, while reducing training and inference costs relative to traditional fine-tuning and prompting techniques.

1837. ST-GCond: Self-supervised and Transferable Graph Dataset Condensation

链接：https://iclr.cc/virtual/2025/poster/27835 abstract： The increasing scale of graph datasets significantly enhances deep learning models but also presents substantial training challenges. Graph dataset condensation has emerged to condense large datasets into smaller yet informative ones that maintain similar test performance. However, these methods require downstream usage to match the original dataset and task, which is impractical in real-world scenarios. Our empirical studies show that existing methods fail in "cross-task" and "cross-dataset" scenarios, often performing worse than training from scratch. To address these challenges, we propose a novel method: Self-supervised and Transferable Graph dataset Condensation (ST-GCond). For cross-task transferability, we propose a task-disentangled meta optimization strategy to adaptively update the condensed graph according to the task relevance, encouraging information preservation for various tasks. For cross-dataset transferability, we propose a multi-teacher self-supervised optimization strategy to incorporate auxiliary self-supervised tasks to inject universal knowledge into the condensed graph. Additionally, we incorporate mutual information guided joint condensation mitigating the potential conflicts and ensure the condensing stability. Experiments on both node-level and graph-level datasets show that ST-GCond outperforms existing methods by 2.5% to 18.7% in all cross-task and cross-dataset scenarios, and also achieves state-of-the-art performance on 5 out of 6 datasets in the single dataset and task scenario.

1838. Asynchronous Federated Reinforcement Learning with Policy Gradient Updates: Algorithm Design and Convergence Analysis

链接：https://iclr.cc/virtual/2025/poster/30959 abstract： To improve the efficiency of reinforcement learning (RL), we propose a novel asynchronous federated reinforcement learning (FedRL) framework termed AFedPG, which constructs a global model through collaboration among N agents using policy gradient (PG) updates. To address the challenge of lagged policies in asynchronous settings, we design a delay-adaptive lookahead technique *specifically for FedRL* that can effectively handle heterogeneous arrival times of policy gradients. We analyze the theoretical global convergence bound of AFedPG, and characterize the advantage of the proposed algorithm in terms of both the sample complexity and time complexity. Specifically, our AFedPG method achieves $\mathcal{O}(\frac{{\epsilon}^{-2.5}}{N})$ sample complexity for global convergence at each agent on average. Compared to the single agent setting with $\mathcal{O}(\epsilon^{-2.5})$ sample complexity, it enjoys a linear speedup with respect to the number of agents. Moreover, compared to synchronous FedPG, AFedPG improves the time complexity from $\mathcal{O}(\frac{t_{\max}}{N})$ to $\mathcal{O}({\sum_{i=1}^{N} \frac{1}{t_{i}}}^{-1}$, where t_{i} denotes the time consumption in each iteration at agent i, and $t_{\max}$ is the largest one. The latter complexity $\mathcal{O}({\sum_{i=1}^{N} \frac{1}{t_{i}}}^{-1}$ is always smaller than the former one, and this improvement becomes significant in large-scale federated settings with heterogeneous computing powers ($t_{\max}\gg t_{\min}$). Finally, we empirically verify the improved performance of AFedPG in four widely used MuJoCo environments with varying numbers of agents. We also demonstrate the advantages of AFedPG in various computing heterogeneity scenarios.

1839. A new framework for evaluating model out-of-distribution generalisation for the biochemical domain

链接：https://iclr.cc/virtual/2025/poster/28260 abstract： Quantifying model generalization to out-of-distribution data has been a longstanding challenge in machine learning. Addressing this issue is crucial for leveraging machine learning in scientific discovery, where models must generalize to new molecules or materials. Current methods typically split data into train and test sets using various criteria — temporal, sequence identity, scaffold, or random cross-validation — before evaluating model performance. However, with so many splitting criteria available, existing approaches offer limited guidance on selecting the most appropriate one, and they do not provide mechanisms for incorporating prior knowledge about the target deployment

distribution(s).To tackle this problem, we have developed a novel metric, AU-GOOD, which quantifies expected model performance under conditions of increasing dissimilarity between train and test sets, while also accounting for prior knowledge about the target deployment distribution(s), when available. This metric is broadly applicable to biochemical entities, including proteins, small molecules, nucleic acids, or cells; as long as a relevant similarity function is defined for them. Recognizing the wide range of similarity functions used in biochemistry, we propose criteria to guide the selection of the most appropriate metric for partitioning. We also introduce a new partitioning algorithm that generates more challenging test sets, and we propose statistical methods for comparing models based on AU-GOOD.Finally, we demonstrate the insights that can be gained from this framework by applying it to two different use cases: developing predictors for pharmaceutical properties of small molecules, and using protein language models as embeddings to build biophysical property predictors.

1840. Think Thrice Before You Act: Progressive Thought Refinement in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28295 abstract： Recent advancements in large language models (LLMs) have demonstrated that progressive refinement, rather than providing a single answer, results in more accurate and thoughtful outputs. However, existing methods often rely heavily on supervision signals to evaluate previous responses, making it difficult to effectively assess output quality in more open-ended scenarios. Additionally, these methods are typically designed for specific tasks, which limits their generalization to new domains. To address these limitations, we propose Progressive Thought Refinement (PTR), a framework that enables LLMs to progressively refine their responses. PTR operates in two phases: (1) Thought data construction stage: We propose a weak and strong model collaborative selection strategy to build a high-quality progressive refinement dataset to ensure logical consistency from thought to answers, and the answers are gradually refined in each round. (2) Thought-Mask Fine-Tuning Phase: We design a training structure to mask the "thought" and adjust loss weights to encourage LLMs to refine prior thought, teaching them to implicitly understand "how to improve" rather than "what is correct." Experimental results show that PTR significantly enhances LLM performance across ten diverse tasks (avg. from 49.6% to 53.5%) without task-specific fine-tuning. Notably, in more open-ended tasks, LLMs also demonstrate substantial improvements in the quality of responses beyond mere accuracy, suggesting that PTR truly teaches LLMs to self-improve over time. Our work is now open-source. https://github.com/cydu24/Progressive-Thought-Refinement

1841. Multiplicative Logit Adjustment Approximates Neural-Collapse-Aware Decision Boundary Adjustment

链接：https://iclr.cc/virtual/2025/poster/30178 abstract： Real-world data distributions are often highly skewed. This has spurred a growing body of research on long-tailed recognition, aimed at addressing the imbalance in training classification models. Among the methods studied, multiplicative logit adjustment (MLA) stands out as a simple and effective method. What theoretical foundation explains the effectiveness of this heuristic method?We provide a justification for the effectiveness of MLA with the following two-step process. First, we develop a theory that adjusts optimal decision boundaries by estimating feature spread on the basis of neural collapse. Second, we demonstrate that MLA approximates this optimal method. Additionally, through experiments on long-tailed datasets, we illustrate the practical usefulness of MLA under more realistic conditions. We also offer experimental insights to guide the tuning of MLA hyperparameters.

1842. Improving Equivariant Networks with Probabilistic Symmetry Breaking

链接：https://iclr.cc/virtual/2025/poster/29230 abstract： Equivariance encodes known symmetries into neural networks, often enhancing generalization. However, equivariant networks cannot break symmetries: the output of an equivariant network must, by definition, have at least the same self-symmetries as its input. This poses an important problem, both (1) for prediction tasks on domains where self-symmetries are common, and (2) for generative models, which must break symmetries in order to reconstruct from highly symmetric latent spaces. This fundamental limitation can in fact be addressed by considering equivariant conditional distributions, instead of equivariant functions. We therefore present novel theoretical results that establish necessary and sufficient conditions for representing such distributions. Concretely, this representation provides a practical framework for breaking symmetries in any equivariant network via randomized canonicalization. Our method, SymPE (Symmetry-breaking Positional Encodings), admits a simple interpretation in terms of positional encodings. This approach expands the representational power of equivariant networks while retaining the inductive bias of symmetry, which we justify through generalization bounds. Experimental results demonstrate that SymPE significantly improves performance of group-equivariant and graph neural networks across diffusion models for graphs, graph autoencoders, and lattice spin system modeling.

1843. Duoduo CLIP: Efficient 3D Understanding with Multi-View Images

链接：https://iclr.cc/virtual/2025/poster/28703 abstract： We introduce Duoduo CLIP, a model for 3D representation learning that learns shape encodings from multi-view images instead of point-clouds. The choice of multi-view images allows us to leverage 2D priors from off-the-shelf CLIP models to facilitate fine-tuning with 3D data. Our approach not only shows better generalization compared to existing point cloud methods, but also reduces GPU requirements and training time. In addition, the model is modified with cross-view attention to leverage information across multiple frames of the object which further boosts performance. Notably, our model is permutation invariant to the order of multi-view images while being pose-free. Compared to the current SOTA point cloud method that requires 480 A100 hours to train 1 billion model parameters we only require 57 A5000 hours and 87 million parameters. Multi-view images also provide more flexibility including being able to encode objects with a

variable number of images, and performance scales when more views are used. In contrast, point cloud based methods require an entire scan or model of the object. We showcase this flexibility with benchmarks from images of real-world objects. Our model also achieves better performance in more fine-grained text to shape retrieval, demonstrating better text-and-shape alignment than point cloud based models.

1844. Conformal Structured Prediction

链接：https://iclr.cc/virtual/2025/poster/31156 abstract： Conformal prediction has recently emerged as a promising strategy for quantifying the uncertainty of a predictive model; these algorithms modify the model to output sets of labels that are guaranteed to contain the true label with high probability. However, existing conformal prediction algorithms have largely targeted classification and regression settings, where the structure of the prediction set has a simple form as a level set of the scoring function. However, for complex structured outputs such as text generation, these prediction sets might include a large number of labels and therefore be hard for users to interpret. In this paper, we propose a general framework for conformal prediction in the structured prediction setting, that modifies existing conformal prediction algorithms to output structured prediction sets that implicitly represent sets of labels. In addition, we demonstrate how our approach can be applied in domains where the prediction sets can be represented as a set of nodes in a directed acyclic graph; for instance, for hierarchical labels such as image classification, a prediction set might be a small subset of coarse labels implicitly representing the prediction set of all their more fine-descendants. We demonstrate how our algorithm can be used to construct prediction sets that satisfy a desired coverage guarantee in several domains.

1845. Straightness of Rectified Flow: A Theoretical Insight into Wasserstein Convergence

链接：https://iclr.cc/virtual/2025/poster/29078 abstract： Diffusion models have emerged as a powerful tool for image generation and denoising. Typically, generative models learn a trajectory between the starting noise distribution and the target data distribution. Recently Liu et al. (2023b) designed a novel alternative generative model Rectified Flow(RF), which aims to learn straight flow trajectories from noise to data using a sequence of convex optimization problems with close ties to optimal transport. If the trajectory is curved, one must use many Euler discretization steps or novel strategies, such as exponential integrators, to achieve a satisfactory generation quality. In contrast, RF has been shown to theoretically straighten the trajectory through successive rectifications, reducing the number of function evaluations (NFEs) while sampling. It has also been shown empirically that RF may improve the straightness in two rectifications if one can solve the underlying optimization problem within a sufficiently small error. In this paper, we make two key theoretical contributions: 1) we provide the first theoretical analysis of theWasserstein distance between the sampling distribution of RF and the target distribution. Our error rate is characterized by the number of discretization steps and a new formulation of straightness stronger than that in the original work. 2) under a mild regularity assumption, we show that for a rectified flow from a Gaussian to any general target distribution with finite first moment (e.g. mixture of Gaussians), two rectifications are sufficient to achieve a straight flow, which is in line with the previous empirical findings. Additionally, we also present empirical results on both simulated and real datasets to validate our theoretical findings.

1846. Rare event modeling with self-regularized normalizing flows: what can we learn from a single failure?

链接：https://iclr.cc/virtual/2025/poster/28821 abstract： Increased deployment of autonomous systems in fields like transportation and robotics have seen a corresponding increase in safety-critical failures. These failures can be difficult to model and debug due to the relative lack of data: compared to tens of thousands of examples from normal operations, we may have only seconds of data leading up to the failure. This scarcity makes it challenging to train generative models of rare failure events, as existing methods risk either overfitting to noise in the limited failure dataset or underfitting due to an overly strong prior. We address this challenge with CalNF, or calibrated normalizing flows, a self-regularized framework for posterior learning from limited data. CalNF achieves state-of-the-art performance on data-limited failure modeling and inverse problems and enables a first-of-a-kind case study into the root causes of the 2022 Southwest Airlines scheduling crisis.

1847. Adaptive Pruning of Pretrained Transformer via Differential Inclusions

链接：https://iclr.cc/virtual/2025/poster/29380 abstract： Large transformers have demonstrated remarkable success, making it necessary to compress these models to reduce inference costs while preserving their performance. Current compression algorithms prune transformers at fixed compression ratios, requiring a unique pruning process for each ratio, which results in high computational costs. In contrast, we propose pruning of pretrained transformers at any desired ratio within a single pruning stage, based on a differential inclusion for a mask parameter. This dynamic can generate the whole regularization solution path of the mask parameter, whose support set identifies the network structure. Therefore, the solution path identifies a Transformer weight family with various sparsity levels, offering greater flexibility and customization.In this paper, weintroduce such an effective pruning method, termed SPP (Solution Path Pruning). To achieve effective pruning, we segment the transformers into paired modules, including query-key pairs, value-projection pairs, and sequential linear layers, and apply low-rank compression to these pairs, maintaining the output structure while enabling structural compression within the inner states. Extensive experiments conducted on various well-known transformer backbones have demonstrated the efficacy of SPP.

1848. Optimal Non-Asymptotic Rates of Value Iteration for Average-Reward Markov Decision Processes

链接：https://iclr.cc/virtual/2025/poster/29337 abstract： While there is an extensive body of research on the analysis of Value Iteration (VI) for discounted cumulative-reward MDPs, prior work on analyzing VI for (undiscounted) average-reward MDPs has been limited, and most prior results focus on asymptotic rates in terms of Bellman error. In this work, we conduct refined non-asymptotic analyses of average-reward MDPs, obtaining a collection of convergence results advancing our understanding of the setup. Among our new results, most notable are the $\mathcal{O}(1/k)$-rates of Anchored Value Iteration on the Bellman error under the multichain setup and the span-based complexity lower bound that matches the $\mathcal{O}(1/k)$ upper bound up to a constant factor of 8 in the weakly communicating and unichain setups.

1849. SmartRAG: Jointly Learn RAG-Related Tasks From the Environment Feedback

链接：https://iclr.cc/virtual/2025/poster/29838 abstract： RAG systems consist of multiple modules to work together. However, these modules are usually separately trained. We argue that a system like RAG that incorporates multiple modules should be jointly optimized to achieve optimal performance. To demonstrate this, we design a specific pipeline called SmartRAG that includes a policy network and a retriever. The policy network can serve as 1) a decision maker that decides when to retrieve, 2) a query rewriter to generate a query most suited to the retriever and 3) an answer generator that produces the final response with/without the observations. We then propose to jointly optimize the whole system using a reinforcement learning algorithm, with the reward designed to encourage the system to achieve the highest performance with minimal retrieval cost. When jointly optimized, each module can be aware of how other modules are working and thus find the best way to work together as a complete system. Empirical results demonstrate that the jointly optimized system can achieve better performance than separately optimized counterparts.

1850. SOREL: A Stochastic Algorithm for Spectral Risks Minimization

链接：https://iclr.cc/virtual/2025/poster/28287 abstract： The spectral risk has wide applications in machine learning, especially in real-world decision-making, where people are concerned with more than just average model performance. By assigning different weights to the losses of different sample points, rather than the same weights as in the empirical risk, it allows the model's performance to lie between the average performance and the worst-case performance. In this paper, we propose SOREL, the first stochastic gradient-based algorithm with convergence guarantees for spectral risks minimization. Previous approaches often rely on smoothing the spectral risk by adding a strongly concave function, thereby lacking convergence guarantees for the original spectral risk. We theoretically prove that our algorithm achieves a near-optimal rate of $\widetilde{O}(1/\sqrt{\epsilon})$ to obtain an ϵ-optimal solution in terms ϵ. Experiments on real datasets show that our algorithm outperforms existing ones in most cases, both in terms of runtime and sample complexity.

1851. Controlling Language and Diffusion Models by Transporting Activations

链接：https://iclr.cc/virtual/2025/poster/28539 abstract： The increasing capabilities of large generative models and their ever more widespread deployment have raised concerns about their reliability, safety, and potential misuse. To address these issues, recent works have proposed to control model generation by steering model activations in order to effectively induce or prevent the emergence of concepts or behaviors in the generated output.In this paper we introduce Activation Transport (AcT), a general framework to steer activations guided by optimal transport theory that generalizes many previous activation-steering works. AcT is modality-agnostic and provides fine-grained control over the model behavior with negligible computational overhead, while minimally impacting model abilities. We experimentally show the effectiveness and versatility of our approach by addressing key challenges in large language models (LLMs) and text-to-image diffusion models (T2Is). For LLMs, we show that AcT can effectively mitigate toxicity, induce arbitrary concepts, and increase their truthfulness. In T2Is, we show how AcT enables fine-grained style control and concept negation.

1852. TopoDiffusionNet: A Topology-aware Diffusion Model

链接：https://iclr.cc/virtual/2025/poster/29223 abstract： Diffusion models excel at creating visually impressive images but often struggle to generate images with a specified topology. The Betti number, which represents the number of structures in an image, is a fundamental measure in topology. Yet, diffusion models fail to satisfy even this basic constraint. This limitation restricts their utility in applications requiring exact control, like robotics and environmental modeling. To address this, we propose TopoDiffusionNet (TDN), a novel approach that enforces diffusion models to maintain the desired topology. We leverage tools from topological data analysis, particularly persistent homology, to extract the topological structures within an image. We then design a topology-based objective function to guide the denoising process, preserving intended structures while suppressing noisy ones. Our experiments across four datasets demonstrate significant improvements in topological accuracy. TDN is the first to integrate topology with diffusion models, opening new avenues of research in this area.

1853. Cheating Automatic LLM Benchmarks: Null Models Achieve High Win Rates

链接：https://iclr.cc/virtual/2025/poster/28083 abstract： Automatic LLM benchmarks, such as AlpacaEval 2.0, Arena-Hard-Auto, and MT-Bench, have become popular for evaluating language models due to their cost-effectiveness and scalability compared to human evaluation. Achieving high win rates on these benchmarks can significantly boost the promotional impact of newly released language models. This promotional benefit may motivate tricks, such as manipulating model output length or style to game win rates, even though several mechanisms have been developed to control length and disentangle style to reduce gameability. Nonetheless, we show that even a **"null model"** that always outputs a **constant** response (*irrelevant to input instructions*) can cheat automatic benchmarks and achieve top-ranked win rates: an 86.5% LC win rate on AlpacaEval 2.0; an 83.0 score on Arena-Hard-Auto; and a 9.55 score on MT-Bench. Moreover, the crafted cheating outputs are **transferable** because we assume that the instructions of these benchmarks (e.g., 805 samples of AlpacaEval 2.0) are *private* and cannot be accessed. While our experiments are primarily proof-of-concept, an adversary could use LLMs to generate more imperceptible cheating responses, unethically benefiting from high win rates and promotional impact. Our findings call for the development of anti-cheating mechanisms for reliable automatic benchmarks. The code is available at https://github.com/sail-sg/Cheating-LLM-Benchmarks.

1854. FairDen: Fair Density-Based Clustering

链接：https://iclr.cc/virtual/2025/poster/29171 abstract： Fairness in data mining tasks like clustering has recently become an increasingly important aspect. However, few clustering algorithms exist that focus on fair groupings of data with sensitive attributes. Including fairness in the clustering objective is especially hard for density-based clustering, as it does not directly optimize a closed form objective like centroid-based or spectral methods. This paper introduces FairDen, the first fair, density-based clustering algorithm.We capture the dataset's density-connectivity structure in a similarity matrix that we manipulate to encourage a balanced clustering. In contrast to state-of-the-art, FairDen inherently handles categorical attributes, noise, and data with several sensitive attributes or groups.We show that FairDen finds meaningful and fair clusters in extensive experiments.

1855. Diffusion Models are Evolutionary Algorithms

链接：https://iclr.cc/virtual/2025/poster/27778 abstract： In a convergence of machine learning and biology, we reveal that diffusion models are evolutionary algorithms. By considering evolution as a denoising process and reversed evolution as diffusion, we mathematically demonstrate that diffusion models inherently perform evolutionary algorithms, naturally encompassing selection, mutation, and reproductive isolation. Building on this equivalence, we propose the Diffusion Evolution method: an evolutionary algorithm utilizing iterative denoising -- as originally introduced in the context of diffusion models -- to heuristically refine solutions in parameter spaces. Unlike traditional approaches, Diffusion Evolution efficiently identifies multiple optimal solutions and outperforms prominent mainstream evolutionary algorithms. Furthermore, leveraging advanced concepts from diffusion models, namely latent space diffusion and accelerated sampling, we introduce Latent Space Diffusion Evolution, which finds solutions for evolutionary tasks in high-dimensional complex parameter space while significantly reducing computational steps. This parallel between diffusion and evolution not only bridges two different fields but also opens new avenues for mutual enhancement, raising questions about open-ended evolution and potentially utilizing non-Gaussian or discrete diffusion models in the context of Diffusion Evolution.

1856. OATS: Outlier-Aware Pruning Through Sparse and Low Rank Decomposition

链接：https://iclr.cc/virtual/2025/poster/30463 abstract： The recent paradigm shift to large-scale foundation models has brought about a new era for deep learning that, while has found great success in practice, has also been plagued by prohibitively expensive costs in terms of high memory consumption and compute. To mitigate these issues, there has been a concerted effort in post-hoc neural network pruning techniques that do not require costly retraining. Despite the considerable progress being made, existing methods often exhibit a steady drop in model performance as the compression increases. In this paper, we present a novel approach to compressing large transformers, coined OATS, that compresses the model weights by approximating each weight matrix as the sum of a sparse matrix and a low-rank matrix. Prior to the decomposition, the weights are first scaled by the second moment of their input embeddings, so as to ensure the preservation of outlier features recently observed in large transformer models. Without retraining, OATS achieves state-of-the-art performance when compressing large language models, such as Llama-3 and Phi-3, and vision transformers, such as Google's ViT and DINOv2, by up to 60%, all while speeding up the model's inference on a CPU by up to $1.37\times$ compared to prior pruning methods.

1857. Discrete Copula Diffusion

链接：https://iclr.cc/virtual/2025/poster/30336 abstract： Discrete diffusion models have recently shown significant progress in modeling complex data, such as natural languages and DNA sequences. However, unlike diffusion models for continuous data, which can generate high-quality samples in just a few denoising steps, modern discrete diffusion models still require hundreds or even thousands of denoising steps to perform well. In this paper, we identify a fundamental limitation that prevents discrete

diffusion models from achieving strong performance with fewer steps -- they fail to capture dependencies between output variables at each denoising step. To address this issue, we provide a formal explanation and introduce a general approach to supplement the missing dependency information by incorporating another deep generative model, termed the copula model. Our method does not require fine-tuning either the diffusion model or the copula model, yet it enables high-quality sample generation with significantly fewer denoising steps. When we apply this approach to autoregressive copula models, the combined model outperforms both models individually in unconditional and conditional text generation. Specifically, the hybrid model achieves better (un)conditional text generation using 8 to 32 times fewer denoising steps than the diffusion model alone. In addition to presenting an effective discrete diffusion generation algorithm, this paper emphasizes the importance of modeling inter-variable dependencies in discrete diffusion.

1858. RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment

链接：https://iclr.cc/virtual/2025/poster/31508 abstract： Generative foundation models are susceptible to implicit biases that can arise from extensive unsupervised training data. Such biases can produce suboptimal samples, skewed outcomes, and unfairness, with potentially serious consequences. Consequently, aligning these models with human ethics and preferences is an essential step toward ensuring their responsible and effective deployment in real-world applications. Prior research has primarily employed Reinforcement Learning from Human Feedback (RLHF) to address this problem, where generative models are fine-tuned with RL algorithms guided by a human-feedback-informed reward model. However, the inefficiencies and instabilities associated with RL algorithms frequently present substantial obstacles to the successful alignment, necessitating the development of a more robust and streamlined approach. To this end, we introduce a new framework, Reward rAnked FineTuning (RAFT), designed to align generative models effectively. Utilizing a reward model and a sufficient number of samples, our approach selects the high-quality samples, discarding those that exhibit undesired behavior, and subsequently enhancing the model by fine-tuning on these filtered samples. Our studies show that RAFT can effectively improve the model performance in both reward learning and other automated metrics in both large language models and diffusion models.

1859. Learn hybrid prototypes for multivariate time series anomaly detection

链接：https://iclr.cc/virtual/2025/poster/30767 abstract： In multivariate time series anomaly detection (MTSAD), reconstruction-based models reconstruct testing series with learned knowledge of only normal series and identify anomalies with higher reconstruction errors. In practice, over-generalization often occurs with unexpectedly well reconstruction of anomalies. Although memory banks are employed by reconstruction-based models to fight against over-generalization, these models are only efficient to detect point anomalies since they learn normal prototypes from time points, leaving interval anomalies and periodical anomalies to be discovered. To settle this problem, this paper propose a hybrid prototypes learning model for MTSAD based on reconstruction, named as H-PAD. First, normal prototypes are learned from different sizes of the patches for time series to discover interval anomalies. These prototypes in different sizes are integrated together to reconstruct query series so that any anomalies would be smoothed off and high reconstruction errors are produced. Furthermore, period prototypes are learned to discover periodical anomalies. One period prototype is memorized for one variable of the query series. Finally, extensive experiments on five benchmark datasets show the effectiveness of H-PAD.

1860. Efficient Biological Data Acquisition through Inference Set Design

链接：https://iclr.cc/virtual/2025/poster/28817 abstract： In drug discovery, highly automated high-throughput laboratories are used to screen a large number of compounds in search of effective drugs. These experiments are expensive, so one might hope to reduce their cost by only experimenting on a subset of the compounds, and predicting the outcomes of the remaining experiments. In this work, we model this scenario as a sequential subset selection problem: we aim to select the smallest set of candidates in order to achieve some desired level of accuracy for the system as a whole. Our key observation is that, if there is heterogeneity in the difficulty of the prediction problem across the input space, selectively obtaining the labels for the hardest examples in the acquisition pool will leave only the relatively easy examples to remain in the inference set, leading to better overall system performance. We call this mechanism inference set design, and propose the use of a confidence-based active learning solution to prune out these challenging examples. Our algorithm includes an explicit stopping criterion that interrupts the acquisition loop when it is sufficiently confident that the system has reached the target performance. Our empirical studies on image and molecular datasets, as well as a real-world large-scale biological assay, show that active learning for inference set design leads to significant reduction in experimental cost while retaining high system performance.

1861. MetaOOD: Automatic Selection of OOD Detection Models

链接：https://iclr.cc/virtual/2025/poster/30662 abstract： How can we automatically select an out-of-distribution (OOD) detection model for various underlying tasks? This is crucial for maintaining the reliability of open-world applications by identifying data distribution shifts, particularly in critical domains such as online transactions, autonomous driving, and real-time patient diagnosis. Despite the availability of numerous OOD detection methods, the challenge of selecting an optimal model for diverse tasks remains largely underexplored, especially in scenarios lacking ground truth labels. In this work, we introduce MetaOOD, the first zero-shot, unsupervised framework that utilizes meta-learning to select an OOD detection model automatically. As a meta-learning approach, MetaOOD leverages historical performance data of existing methods across various benchmark OOD detection datasets, enabling the effective selection of a suitable model for new datasets without the need for labeled data at the test time. To quantify task similarities more accurately, we introduce language model-based embeddings that capture the

distinctive OOD characteristics of both datasets and detection models. Through extensive experimentation with 24 unique test dataset pairs to choose from among 11 OOD detection models, we demonstrate that MetaOOD significantly outperforms existing methods and only brings marginal time overhead. Our results, validated by Wilcoxon statistical tests, show that MetaOOD surpasses a diverse group of 11 baselines, including established OOD detectors and advanced unsupervised selection methods.

1862. Transformer Block Coupling and its Correlation with Generalization in LLMs

链接：https://iclr.cc/virtual/2025/poster/28555 abstract： Large Language Models (LLMs) have made significant strides in natural language processing, and a precise understanding of the internal mechanisms driving their success is essential. In this work, we analyze the trajectories of token embeddings as they pass through transformer blocks, linearizing the system along these trajectories through their Jacobian matrices. By examining the relationships between these block Jacobians, we uncover the phenomenon of transformer block coupling in a multitude of LLMs, characterized by the coupling of their top singular vectors across tokens and depth. Our findings reveal that coupling positively correlates with model performance, and that this relationship is stronger than with other hyperparameters such as parameter count, model depth, and embedding dimension. We further investigate how these properties emerge during training, observing a progressive development of coupling, increased linearity, and layer-wise exponential growth in token trajectories. Additionally, experiments with Vision Transformers (ViTs) corroborate the emergence of coupling and its relationship with generalization, reinforcing our findings in LLMs. Collectively, these insights offer a novel perspective on token interactions in transformers, opening new directions for studying their mechanisms as well as improving training and generalization.

1863. Collapsed Language Models Promote Fairness

链接：https://iclr.cc/virtual/2025/poster/28548 abstract： To mitigate societal biases implicitly encoded in recent successful pretrained language models, a diverse array of approaches have been proposed to encourage model fairness, focusing on prompting, data augmentation, regularized fine-tuning, and more. Despite the development, it is nontrivial to reach a principled understanding of fairness and an effective algorithm that can consistently debias language models. In this work, by rigorous evaluations of Neural Collapse -- a learning phenomenon happen in last-layer representations and classifiers in deep networks -- on fairness-related words, we find that debiased language models exhibit collapsed alignment between token representations and word embeddings. More importantly, this observation inspires us to design a principled fine-tuning method that can effectively improve fairness in a wide range of debiasing methods, while still preserving the performance of language models on standard natural language understanding tasks. We attach our code at https://github.com/Xujxyang/Fairness-NC-main

1864. Glauber Generative Model: Discrete Diffusion Models via Binary Classification

链接：https://iclr.cc/virtual/2025/poster/30199 abstract： We introduce the Glauber Generative Model (GGM), a new class of discrete diffusion models, to obtain new samples from a distribution given samples from a discrete space. GGM deploys a discrete Markov chain called the heat bath dynamics (or the Glauber dynamics) to denoise a sequence of noisy tokens to a sample from a joint distribution of discrete tokens. Our novel conceptual framework provides an exact reduction of the task of learning the denoising Markov chain to solving a class of binary classification tasks. More specifically, the model learns to classify a given token in a noisy sequence as signal or noise. In contrast, prior works on discrete diffusion models either solve regression problems to learn importance ratios, or minimize loss functions given by variational approximations. We apply GGM to language modeling and image generation, where images are discretized using image tokenizers like VQGANs. We show that it outperforms existing discrete diffusion models in language generation, and demonstrates strong performance for image generation without using dataset-specific image tokenizers. We also show that our model is capable of performing well in zero-shot control settings like text and image infilling.

1865. SMITE: Segment Me In TimE

链接：https://iclr.cc/virtual/2025/poster/30055 abstract： Segmenting an object in a video presents significant challenges. Each pixel must be accurately labeled, and these labels must remain consistent across frames. The difficulty increases when the segmentation is with arbitrary granularity, meaning the number of segments can vary arbitrarily, and masks are defined based on only one or a few sample images. In this paper, we address this issue by employing apre-trained text to image diffusion model supplemented with an additional tracking mechanism. We demonstrate that our approach can effectively manage various segmentation scenarios and outperforms state-of-the-art alternatives. The project page is available at https://segment-me-in-time.github.io/

1866. Privacy-Preserving Personalized Federated Prompt Learning for Multimodal Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30385 abstract： Multimodal Large Language Models (LLMs) are pivotal in revolutionizing customer support and operations by integrating multiple modalities such as text, images, and audio. Federated

Prompt Learning (FPL) is a recently proposed approach that combines pre-trained multimodal LLMs such as vision-language models with federated learning to create personalized, privacy-preserving AI systems. However, balancing the competing goals of personalization, generalization, and privacy remains a significant challenge. Over-personalization can lead to overfitting, reducing generalizability, while stringent privacy measures, such as differential privacy, can hinder both personalization and generalization. In this paper, we propose a Differentially Private Federated Prompt Learning (DP-FPL) approach to tackle this challenge by leveraging a low-rank factorization scheme to capture generalization while maintaining a residual term that preserves expressiveness for personalization. To ensure privacy, we introduce a novel method where we apply local differential privacy to the two low-rank components of the local prompt, and global differential privacy to the global prompt. Our approach mitigates the impact of privacy noise on the model performance while balancing the tradeoff between personalization and generalization. Extensive experiments demonstrate the effectiveness of our approach over other benchmarks.

1867. LLaMaFlex: Many-in-one LLMs via Generalized Pruning and Weight Sharing

链接：https://iclr.cc/virtual/2025/poster/30601 abstract： Large Language Model (LLM) providers typically train a family of models, each of a different size targeting a specific deployment scenario. Models in the family are all trained from scratch, making the process extremely resource intensive.Recent work has successfully reduced the cost of training model families through a combination of structured pruning and knowledge distillation; here, only the largest model in the family is trained from scratch, and smaller models are obtained via pruning. We observe that while effective, this strategy must still perform pruning and distillation with hundreds of billions of training tokens for every new model, keeping overall training costs high.In this work, we introduce a novel nested weight-shared architecture named LLaMaFlex that can be pruned across both width and depth dimensions in a zero-shot manner to instantly yield a large number of highly accurate compressed models.LLaMaFlex starts from a pretrained model, and only requires a single continued training phase consisting of ~60B tokens, which trains the elastic network and an end-to-end Gumbel Softmax-based router; this router is able to interpolate smoothly across model sizes, enabling the "train once, deploy many" paradigm.We train LLaMaFlex on Llama 3.1 8B and use it to zero-shot generate a family of compressed models that achieves accuracy on par with or better than state-of-the-art pruned, elastic/flexible, and trained-from-scratch models.

1868. Once-for-All: Controllable Generative Image Compression with Dynamic Granularity Adaptation

链接：https://iclr.cc/virtual/2025/poster/27695 abstract： Although recent generative image compression methods have demonstrated impressive potential in optimizing the rate-distortion-perception trade-off, they still face the critical challenge of flexible rate adaptation to diverse compression necessities and scenarios. To overcome this challenge, this paper proposes a $\textbf{Control}$lable $\textbf{G}$enerative $\textbf{I}$mage $\textbf{C}$ompression framework, $\textbf{Control-GIC}$, the first capable of fine-grained bitrate adaptation across a broad spectrum while ensuring high-fidelity and generality compression. We base Control-GIC on a VQGAN framework representing an image as a sequence of variable-length codes ($\textit{i.e.}$ VQ-indices), which can be losslessly compressed and exhibits a direct positive correlation with bitrates. Drawing inspiration from the classical coding principle, we correlate the information density of local image patches with their granular representations. Hence, we can flexibly determine a proper allocation of granularity for the patches to achieve dynamic adjustment for VQ-indices, resulting in desirable compression rates. We further develop a probabilistic conditional decoder capable of retrieving historic encoded multi-granularity representations according to transmitted codes, and then reconstruct hierarchical granular features in the formalization of conditional probability, enabling more informative aggregation to improve reconstruction realism. Our experiments show that Control-GIC allows highly flexible and controllable bitrate adaptation where the results demonstrate its superior performance over recent state-of-the-art methods.

1869. Generative Monoculture in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/27717 abstract： We introduce {\em generative monoculture}, a behavior observed in large language models (LLMs) characterized by a significant narrowing of model output diversity relative to available training data for a given task: for example, generating only positive book reviews for books with a mixed reception. While in some cases, generative monoculture enhances performance (e.g., LLMs more often produce efficient code), the dangers are exacerbated in others (e.g., LLMs refuse to share diverse opinions). As LLMs are increasingly used in high-impact settings such as education and web search, careful maintenance of LLM output diversity is essential to ensure a variety of facts and perspectives are preserved over time. We experimentally demonstrate the prevalence of generative monoculture through analysis of book review and code generation tasks, and find that simple countermeasures such as altering sampling or prompting strategies are insufficient to mitigate the behavior. Moreover, our results suggest that the root causes of generative monoculture are likely embedded within the LLM's alignment processes, suggesting a need for developing fine-tuning paradigms that preserve or promote diversity.

1870. ClassDiffusion: More Aligned Personalization Tuning with Explicit Class Guidance

链接：https://iclr.cc/virtual/2025/poster/28697 abstract：

1871. Image and Video Tokenization with Binary Spherical Quantization

链接：https://iclr.cc/virtual/2025/poster/27738 abstract： We propose a new transformer-based image and video tokenizer with Binary Spherical Quantization (BSQ). BSQ projects the high-dimensional visual embedding to a lower-dimensional hypersphere and then applies binary quantization. BSQ is (1) parameter-efficient without an explicit codebook, (2) scalable to arbitrary token dimensions, and (3) compact: compressing visual data by up to 100×with minimal distortion. Our tokenizer uses a transformer encoder and decoder with simple block-wise causal masking to support variable-length videos as input. The resulting BSQ-ViT achieves state-of-the-art visual reconstruction quality on image and video reconstruction benchmarks with 2.4× throughput compared to the best prior methods. Furthermore, by learning an autoregressive prior for adaptive arithmetic coding, BSQ-ViT achieves comparable visual compression results with commonly used compression standards, e.g. JPEG2000/WebP for images and H.264/H.265 for videos. BSQ-ViT also enables masked language models to achieve competitive image synthesis quality to GAN and diffusion approaches.

1872. Training Robust Ensembles Requires Rethinking Lipschitz Continuity

链接：https://iclr.cc/virtual/2025/poster/29370 abstract：

1873. MonST3R: A Simple Approach for Estimating Geometry in the Presence of Motion

链接：https://iclr.cc/virtual/2025/poster/28524 abstract：

1874. Distilling Structural Representations into Protein Sequence Models

链接：https://iclr.cc/virtual/2025/poster/30051 abstract： Protein language (or sequence) models, like the popular ESM2, are now widely used tools for extracting evolution-based protein representations and have achieved significant success on core downstream biological tasks.A major open problem is how to obtain representations that best capture both the sequence evolutionary history and the atomic structural properties of proteins in general. We introduce Implicit Sequence Model, a sequence-only input model with structurally-enriched representations that outperforms state-of-the-art sequence models on several well-studied benchmarks including mutation stability assessment and structure prediction. Our key innovations are a microenvironment-based Autoencoder for generating structure tokens and a self-supervised training objective that distills these tokens into ESM2's pre-trained model. Notably, we make ISM's structure-enriched weights easily accessible for any application using the ESM2 framework.

1875. Motion-Agent: A Conversational Framework for Human Motion Generation with LLMs

链接：https://iclr.cc/virtual/2025/poster/30605 abstract： While previous approaches to 3D human motion generation have achieved notable success, they often rely on extensive training and are limited to specific tasks. To address these challenges, we introduce Motion-Agent, an efficient conversational framework designed for general human motion generation, editing, and understanding. Motion-Agent employs an open-source pre-trained language model to develop a generative agent, MotionLLM, that bridges the gap between motion and text. This is accomplished by encoding and quantizing motions into discrete tokens that align with the language model's vocabulary. With only 1-3% of the model's parameters fine-tuned using adapters, MotionLLM delivers performance on par with diffusion models and other transformer-based methods trained from scratch. By integrating MotionLLM with GPT-4 without additional training, Motion-Agent is able to generate highly complex motion sequences through multi-turn conversations, a capability that previous models have struggled to achieve.Motion-Agent supports a wide range of motion-language tasks, offering versatile capabilities for generating and customizing human motion through interactive conversational exchanges.

1876. COPER: Correlation-based Permutations for Multi-View Clustering

链接：https://iclr.cc/virtual/2025/poster/30937 abstract： Combining data from different sources can improve data analysis tasks such as clustering. However, most of the current multi-view clustering methods are limited to specific domains or rely on a suboptimal and computationally intensive two-stage process of representation learning and clustering. We propose an end-to-end deep learning-based multi-view clustering framework for general data types (such as images and tables). Our approach involves generating meaningful fused representations using a novel permutation-based canonical correlation objective. We provide a theoretical analysis showing how the learned embeddings approximate those obtained by supervised linear discriminant analysis (LDA). Cluster assignments are learned by identifying consistent pseudo-labels across multiple views. Additionally, we establish a theoretical bound on the error caused by incorrect pseudo-labels in the unsupervised representations compared to LDA. Extensive experiments on ten multi-view clustering benchmark datasets provide empirical evidence for the effectiveness of the proposed model.

1877. MagicDec: Breaking the Latency-Throughput Tradeoff for Long

Context Generation with Speculative Decoding

链接：https://iclr.cc/virtual/2025/poster/30513 abstract： Large Language Models (LLMs) have become more prevalent in long-context applications such as interactive chatbots, document analysis, and agent workflows, but it is challenging to serve long-context requests with low latency and high throughput. Speculative decoding (SD) is a widely used technique to reduce latency losslessly, but the conventional wisdom suggests that its efficacy is limited to small batch sizes. In MagicDec, we show that surprisingly SD can achieve speedup even for a high throughput inference regime for moderate to long sequences. More interestingly, an intelligent drafting strategy can achieve better speedup with increasing batch size based on our rigorous analysis. MagicDec first identifies the bottleneck shifts with increasing batch size and sequence length, and uses these insights to deploy SD more effectively for high throughput inference. We leverage draft model with sparse KV cache to address the KV bottleneck, which scales with both sequence length and batch size. Additionally, we propose a theoretical model to select the optimal drafting strategy for maximum speedup. Our work highlights the broad applicability of speculative decoding in long-context serving, as it can enhance throughput and reduce latency without compromising accuracy. For moderate to long sequences, we demonstrate up to 2.51x speedup for LLaMA-3.1-8B when serving batch sizes ranging from 32 to 256 on various types of hardware and tasks.

1878. KinPFN: Bayesian Approximation of RNA Folding Kinetics using Prior-Data Fitted Networks

链接：https://iclr.cc/virtual/2025/poster/30428 abstract： RNA is a dynamic biomolecule crucial for cellular regulation, with its function largely determined by its folding into complex structures, while misfolding can lead to multifaceted biological sequelae. During the folding process, RNA traverses through a series of intermediate structural states, with each transition occurring at variable rates that collectively influence the time required to reach the functional form. Understanding these folding kinetics is vital for predicting RNA behavior and optimizing applications in synthetic biology and drug discovery. While in silico kinetic RNA folding simulators are often computationally intensive and time-consuming, accurate approximations of the folding times can already be very informative to assess the efficiency of the folding process. In this work, we present KinPFN, a novel approach that leverages prior-data fitted networks to directly model the posterior predictive distribution of RNA folding times. By training on synthetic data representing arbitrary prior folding times, KinPFN efficiently approximates the cumulative distribution function of RNA folding times in a single forward pass, given only a few initial folding time examples. Our method offers a modular extension to existing RNA kinetics algorithms, promising significant computational speed-ups orders of magnitude faster, while achieving comparable results. We showcase the effectiveness of KinPFN through extensive evaluations and real-world case studies, demonstrating its potential for RNA folding kinetics analysis, its practical relevance, and generalization to other biological data.

1879. SMT: Fine-Tuning Large Language Models with Sparse Matrices

链接：https://iclr.cc/virtual/2025/poster/30274 abstract： Various parameter-efficient fine-tuning (PEFT) methods, including LoRA and its variants, have gained popularity for reducing computational costs. However, there is often an accuracy gap between PEFT approaches and full fine-tuning (FT), and this discrepancy has not yet been systematically explored. In this work, we introduce a method for selecting sparse sub-matrices that aims to minimize the performance gap between PEFT vs. full fine-tuning (FT) while also reducing both fine-tuning computational costs and memory costs. We explored both gradient-based and activation-based parameter selection methods to identify the most significant sub-matrices for downstream tasks, updating only these blocks during fine-tuning. In our experiments, we demonstrated that SMT consistently surpasses other PEFTbaselines (e.g., LoRA and DoRA) in fine-tuning popular large language models such as LLaMA across a broad spectrum of tasks, while reducing the GPU memory footprint by 67% compared to FT. We also examine how the performance of LoRA and DoRA tends to plateau and decline as the number of trainable parameters increases, in contrast, our SMT method does not suffer from such issues.

1880. Provence: efficient and robust context pruning for retrieval-augmented generation

链接：https://iclr.cc/virtual/2025/poster/29557 abstract： Retrieval-Augmented Generation improves various aspects of large language models (LLMs) generation, but suffers from computational overhead caused by long contexts, and the propagation of irrelevant retrieved information into generated responses. Context pruning deals with both aspects, by removing irrelevant parts of retrieved contexts before LLM generation. Existing context pruning approaches are limited, and do not present a universal model that would be both efficient and robust in a wide range of scenarios, e.g., when contexts contain a variable amount of relevant information or vary in length, or when evaluated on various domains. In this work, we close this gap and introduce Provence (Pruning and Reranking Of retrieVEd relevaNt ContExts), an efficient and robust context pruner for Question Answering, which dynamically detects the needed amount of pruning for a given context and can be used out-of-the-box for various domains. The three key ingredients of Provence are formulating the context pruning task as sequence labeling, unifying context pruning capabilities with context reranking, and training on diverse data. Our experimental results show that Provence enables context pruning with negligible to no drop in performance, in various domains and settings, at almost no cost in a standard RAG pipeline. We also conduct a deeper analysis alongside various ablations to provide insights into training context pruners for future work.

1881. Long-Context Linear System Identification

链接：https://iclr.cc/virtual/2025/poster/31134 abstract： This paper addresses the problem of long-context linear system identification, where the state x_t of the system at time t depends linearly on previous states x_s over a fixed context window of length p. We establish a sample complexity bound that matches the *i.i.d.* parametric rate, up to logarithmic factors for a broad class of systems, extending previous work that considered only first-order dependencies. Our findings reveal a ``learning-without-mixing'' phenomenon, indicating that learning long-context linear autoregressive models is not hindered by slow mixing properties potentially associated with extended context windows. Additionally, we extend these results to *(i)* shared low-rank feature representations, where rank-regularized estimators improve rates with respect to dimensionality, and *(ii)* misspecified context lengths in strictly stable systems, where shorter contexts offer statistical advantages.

1882. Spectro-Riemannian Graph Neural Networks

链接：https://iclr.cc/virtual/2025/poster/31145 abstract： Can integrating spectral and curvature signals unlock new potential in graph representation learning? Non-Euclidean geometries, particularly Riemannian manifolds such as hyperbolic (negative curvature) and spherical (positive curvature), offer powerful inductive biases for embedding complex graph structures like scale-free, hierarchical, and cyclic patterns. Meanwhile, spectral filtering excels at processing signal variations across graphs, making it effective in homophilic and heterophilic settings. Leveraging both can significantly enhance the learned representations. To this end, we propose Spectro-Riemannian Graph Neural Networks (CUSP) - the first graph representation learning paradigm that unifies both CUrvature (geometric) and SPectral insights. CUSP is a mixed-curvature spectral GNN that learns spectral filters to optimize node embeddings in products of constant curvature manifolds (hyperbolic, spherical, and Euclidean). Specifically, CUSP introduces three novel components: (a) Cusp Laplacian, an extension of the traditional graph Laplacian based on Ollivier-Ricci curvature, designed to capture the curvature signals better; (b) Cusp Filtering, which employs multiple Riemannian graph filters to obtain cues from various bands in the eigenspectrum; and (c) Cusp Pooling, a hierarchical attention mechanism combined with a curvature-based positional encoding to assess the relative importance of differently curved substructures in our graph. Empirical evaluation across eight homophilic and heterophilic datasets demonstrates the superiority of CUSP in node classification and link prediction tasks, with a gain of up to 5.3\% over state-of-the-art models.

1883. Efficient Evolutionary Search Over Chemical Space with Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29136 abstract： Molecular discovery, when formulated as an optimization problem, presents significant computational challenges because optimization objectives can be non-differentiable. Evolutionary Algorithms (EAs), often used to optimize black-box objectives in molecular discovery, traverse chemical space by performing random mutations and crossovers, leading to a large number of expensive objective evaluations. In this work, we ameliorate this shortcoming by incorporating chemistry-aware Large Language Models (LLMs) into EAs. Namely, we redesign crossover and mutation operations in EAs using LLMs trained on large corpora of chemical information. We perform extensive empirical studies on both commercial and open-source models on multiple tasks involving property optimization, molecular rediscovery, and structure-based drug design, demonstrating that the joint usage of LLMs with EAs yields superior performance over all baseline models across single- and multi-objective settings. We demonstrate that our algorithm improves both the quality of the final solution and convergence speed, thereby reducing the number of required objective evaluations.

1884. Efficient Sparse PCA via Block-Diagonalization

链接：https://iclr.cc/virtual/2025/poster/30361 abstract： Sparse Principal Component Analysis (Sparse PCA) is a pivotal tool in data analysis and dimensionality reduction. However, Sparse PCA is a challenging problem in both theory and practice: it is known to be NP-hard and current exact methods generally require exponential runtime. In this paper, we propose a novel framework to efficiently approximate Sparse PCA by (i) approximating the general input covariance matrix with a re-sorted block-diagonal matrix, (ii) solving the Sparse PCA sub-problem in each block, and (iii) reconstructing the solution to the original problem. Our framework is simple and powerful: it can leverage any off-the-shelf Sparse PCA algorithm and achieve significant computational speedups, with a minor additive error that is linear in the approximation error of the block-diagonal matrix. Suppose $g(k, d)$ is the runtime of an algorithm (approximately) solving Sparse PCA in dimension d and with sparsity constant k. Our framework, when integrated with this algorithm, reduces the runtime to $\mathcal{O}\left(\frac{d}{d^\star} \cdot g(k, d^\star) + d^2\right)$, where $d^\star \leq d$ is the largest block size of the block-diagonal matrix. For instance, integrating our framework with the Branch-and-Bound algorithm reduces the complexity from $g(k, d) = \mathcal{O}(k^3\cdot d^k)$ to $\mathcal{O}(k^3\cdot d \cdot (d^\star)^{k-1})$, demonstrating exponential speedups if $d^\star$ is small. We perform large-scale evaluations on many real-world datasets: for exact Sparse PCA algorithm, our method achieves an average speedup factor of 100.50, while maintaining an average approximation error of 0.61%; for approximate Sparse PCA algorithm, our method achieves an average speedup factor of 6.00 and an average approximation error of -0.91%, meaning that our method oftentimes finds better solutions.

1885. Chain-of-region: Visual Language Models Need Details for Diagram Analysis

链接：https://iclr.cc/virtual/2025/poster/29954 abstract： Visual Language Models (VLMs) like GPT-4V have broadened the scope of LLM applications, yet they face significant challenges in accurately processing visual details, particularly in scientific diagrams. This paper explores the necessity of meticulous visual detail collection and region decomposition for enhancing the performance of VLMs in scientific diagram analysis. We propose a novel approach that combines traditional computer vision techniques with VLMs to systematically decompose diagrams into discernible visual elements and aggregate essential metadata. Our method employs techniques in OpenCV library to identify and label regions, followed by a refinement process using shape detection and region merging algorithms, which are particularly suited to the structured nature of scientific diagrams. This strategy not only improves the granularity and accuracy of visual information processing but also extends the capabilities of VLMs beyond their current limitations. We validate our approach through a series of experiments that demonstrate enhanced performance in diagram analysis tasks, setting a new standard for integrating visual and language processing in a multimodal context.

1886. Reasoning of Large Language Models over Knowledge Graphs with Super-Relations

链接：https://iclr.cc/virtual/2025/poster/28196 abstract： While large language models (LLMs) have made significant progress in processing and reasoning over knowledge graphs, current methods suffer from a high non-retrieval rate. This limitation reduces the accuracy of answering questions based on these graphs. Our analysis reveals that the combination of greedy search and forward reasoning is a major contributor to this issue. To overcome these challenges, we introduce the concept of super-relations, which enables both forward and backward reasoning by summarizing and connecting various relational paths within the graph. This holistic approach not only expands the search space, but also significantly improves retrieval efficiency. In this paper, we propose the ReKnoS framework, which aims to Reason over Knowledge Graphs with Super-Relations. Our framework's key advantages include the inclusion of multiple relation paths through super-relations, enhanced forward and backward reasoning capabilities, and increased efficiency in querying LLMs. These enhancements collectively lead to a substantial improvement in the successful retrieval rate and overall reasoning performance. We conduct extensive experiments on a variety of datasets to evaluate ReKnoS, and the results demonstrate the superior performance of ReKnoS over existing state-of-the-art baselines, with an average accuracy gain of 2.92% across nine real-world datasets.

1887. Aligned LLMs Are Not Aligned Browser Agents

链接：https://iclr.cc/virtual/2025/poster/29856 abstract： For safety reasons, large language models (LLMs) are trained to refuse harmful user instructions, such as assisting dangerous activities. We study an open question in this work: does the desired safety refusal, typically enforced in chat contexts, generalize to non-chat and agentic use cases? Unlike chatbots, LLM agents equipped with general-purpose tools, such as web browsers and mobile devices, can directly influence the real world, making it even more crucial to refuse harmful instructions. In this work, we primarily focus on red-teaming browseragents – LLMs that leverage information via web browsers. To this end, we introduce Browser Agent Red teaming Toolkit (BrowserART), a comprehensive test suite designed specifically for red-teaming browser agents. BrowserART consists of 100 diverse browser-related harmful behaviors (including original behaviors and ones sourced from HarmBench (Mazeika et al., 2024) and AirBench 2024 (Zeng et al., 2024b)) across both synthetic and real websites. Our empirical study on state-of-the-art browser agents reveals that while the backbone LLM refuses harmful instructions as a chatbot, the corresponding agent does not. Moreover, attack methods designed to jailbreak refusal-trained LLMs in the chat settings transfer effectively to browser agents. With human rewrites, GPT-4o and o1-preview -based browser agents pursued 98 and 63 harmful behaviors (out of 100), respectively. Therefore, simply ensuring LLM's refusal to harmful instruc-tions in chats is not sufficient to ensure that the downstream agents are safe. We publicly release BrowserART and call on LLM developers, policymakers, and agent developers to collaborate on improving agent safety.

1888. SiMHand: Mining Similar Hands for Large-Scale 3D Hand Pose Pre-training

链接：https://iclr.cc/virtual/2025/poster/30721 abstract： We present a framework for pre-training of 3D hand pose estimation from in-the-wild hand images sharing with similar hand characteristics, dubbed SiMHand. Pre-training with large-scale images achieves promising results in various tasks, but prior methods for 3D hand pose pre-training have not fully utilized the potential of diverse hand images accessible from in-the-wild videos. To facilitate scalable pre-training, we first prepare an extensive pool of hand images from in-the-wild videos and design our pre-training method with contrastive learning. Specifically, we collect over 2.0M hand images from recent human-centric videos, such as 100DOH and Ego4D. To extract discriminative information from these images, we focus on the similarity of hands: pairs of non-identical samples with similar hand poses. We then propose a novel contrastive learning method that embeds similar hand pairs closer in the feature space. Our method not only learns from similar samples but also adaptively weights the contrastive learning loss based on inter-sample distance, leading to additional performance gains. Our experiments demonstrate that our method outperforms conventional contrastive learning approaches that produce positive pairs solely from a single image with data augmentation. We achieve significant improvements over the state-of-the-art method (PeCLR) in various datasets, with gains of 15% on FreiHand, 10% on DexYCB, and 4% on AssemblyHands. Our code is available at https://github.com/ut-vision/SiMHand.

1889. Minimal Impact ControlNet: Advancing Multi-ControlNet Integration

链接：https://iclr.cc/virtual/2025/poster/28156 abstract： With the advancement of diffusion models, there is a growing demand for high-quality, controllable image generation, particularly through methods that utilize one or multiple control signals based on ControlNet. However, in current ControlNet training, each control is designed to influence all areas of an image, which can lead to conflicts when different control signals are expected to manage different parts of the image in practical applications. This issue is especially pronounced with edge-type control conditions, where regions lacking boundary information often represent low-frequency signals, referred to as silent control signals. When combining multiple ControlNets, these silent control signals can suppress the generation of textures in related areas, resulting in suboptimal outcomes. To address this problem, we propose Minimal Impact ControlNet. Our approach mitigates conflicts through three key strategies: constructing a balanced dataset, combining and injecting feature signals in a balanced manner, and addressing the asymmetry in the score function’s Jacobian matrix induced by ControlNet. These improvements enhance the compatibility of control signals, allowing for freer and more harmonious generation in areas with silent control signals.

1890. Unbounded: A Generative Infinite Game of Character Life Simulation

链接：https://iclr.cc/virtual/2025/poster/27943 abstract： We introduce the concept of a generative infinite game, a video game that transcends the traditional boundaries of finite, hard-coded systems by using generative models. Inspired by James P. Carse's distinction between finite and infinite games, we leverage recent advances in generative AI to create Unbounded: a game of character life simulation that is fully encapsulated in generative models. Specifically, Unbounded draws inspiration from sandbox life simulations and allows you to interact with your autonomous virtual character in a virtual world by feeding, playing with and guiding it - with open-ended mechanics generated by an LLM, some of which can be emergent. In order to develop Unbounded, we propose technical innovations in both the LLM and visual generation domains. Specifically, we present: (1) a specialized, distilled large language model (LLM) that dynamically generates game mechanics, narratives, and character interactions in real-time, and (2) a new dynamic regional image prompt Adapter (IP-Adapter) for vision models that ensures consistent yet flexible visual generation of a character across multiple environments. We evaluate our system through both qualitative and quantitative analysis, showing significant improvements in character life simulation, user instruction following, narrative coherence, and visual consistency for both characters and the environments compared to traditional related approaches.

1891. Noise Stability Optimization for Finding Flat Minima: A Hessian-based Regularization Approach

链接：https://iclr.cc/virtual/2025/poster/31468 abstract： The training of over-parameterized neural networks has received much study in recent literature. An important consideration is the regularization of over-parameterized networks due to their highly nonconvex and nonlinear geometry. In this paper, we study noise injection algorithms, which can regularize the Hessian of the loss, leading to regions with flat loss surfaces. Specifically, by injecting isotropic Gaussian noise into the weight matrices of a neural network, we can obtain an approximately unbiased estimate of the trace of the Hessian. However, naively implementing the noise injection via adding noise to the weight matrices before backpropagation presents limited empirical improvements. To address this limitation, we design a two-point estimate of the Hessian penalty, which injects noise into the weight matrices along both positive and negative directions of the random noise. In particular, this two-point estimate eliminates the variance of the first-order Taylor's expansion term on the Hessian. We show a PAC-Bayes generalization bound that depends on the trace of the Hessian (and the radius of the weight space), which can be measured from data. We conduct a detailed experimental study to validate our approach and show that it can effectively regularize the Hessian and improve generalization. First, our algorithm can outperform prior approaches on sharpness-reduced training, delivering up to a 2.4% test accuracy increase for fine-tuning ResNets on six image classification datasets. Moreover, the trace of the Hessian reduces by 15.8%, and the largest eigenvalue is reduced by 9.7% with our approach. We also find that the regularization of the Hessian can be combined with alternative regularization methods, such as weight decay and data augmentation, leading to stronger regularization. Second, our approach remains highly effective for improving generalization in pretraining multimodal CLIP models and chain-of-thought fine-tuning.

1892. Three-in-One: Fast and Accurate Transducer for Hybrid-Autoregressive ASR

链接：https://iclr.cc/virtual/2025/poster/29973 abstract： We present Hybrid-Autoregressive INference TrANsducers (HAINAN), a novel architecture for speech recognition that extends the Token-and-Duration Transducer (TDT) model. Trained with randomly masked predictor network outputs, HAINAN supports both autoregressive inference with all network components and non-autoregressive inference without the predictor. Additionally, we propose a novel semi-autoregressive inference method that first generates an initial hypothesis using non-autoregressive inference, followed by refinement steps where each token prediction is regenerated using parallelized autoregression on the initial hypothesis. Experiments on multiple datasets across different languages demonstrate that HAINAN achieves efficiency parity with CTC in non-autoregressive mode and with TDT in autoregressive mode. In terms of accuracy, autoregressive HAINAN achieves parity with TDT and RNN-T, while non-autoregressive HAINAN significantly outperforms CTC. Semi-autoregressive inference further enhances the model's accuracy with minimal computational overhead, and even outperforms TDT results in some cases. These results highlight HAINAN's flexibility in balancing accuracy and speed, positioning it as a strong candidate for real-world speech recognition applications.

1893. Fluid: Scaling Autoregressive Text-to-image Generative Models with

Continuous Tokens

链接：https://iclr.cc/virtual/2025/poster/31519 abstract： Scaling up autoregressive models in vision has not proven as beneficial as in large language models.In this work, we investigate this scaling problem in the context of text-to-image generation, focusing on two critical factors: whether models use discrete or continuous tokens, and whether tokens are generated in a random or fixed raster order using BERT- or GPT-like transformer architectures. Our empirical results show that, while all models scale effectively in terms of validation loss, their evaluation performance -- measured by FID, GenEval score, and visual quality -- follows different trends. Models based on continuous tokens achieves significantly better visual quality than those using discrete tokens. Furthermore, the generation order and attention mechanisms significantly affect the GenEval score: random-order models achieve notably better GenEval scores compared to raster-order models.Inspired by these findings, we train Fluid, a random-order autoregressive model on continuous tokens. Fluid 10.5B model achieves a new state-of-the-art zeor-shot FID of 6.16 on MS-COCO 30K, and 0.69 overall score on the GenEval benchmark. We hope our findings and results will encourage future efforts to further bridge the scaling gap between vision and language models.

1894. A Unified Framework for Forward and Inverse Problems in Subsurface Imaging using Latent Space Translations

链接：https://iclr.cc/virtual/2025/poster/27736 abstract： In subsurface imaging, learning the mapping from velocity maps to seismic waveforms (forward problem) and waveforms to velocity (inverse problem) is important for several applications. While traditional techniques for solving forward and inverse problems are computationally prohibitive, there is a growing interest to leverage recent advances in deep learning to learn the mapping between velocity maps and seismic waveform images directly from data. Despite the variety of architectures explored in previous works, several open questions still remain unanswered such as the effect of latent space sizes, the importance of manifold learning, the complexity of translation models, and the value of jointly solving forward and inverse problems. We propose a unified framework to systematically characterize prior research in this area termed the Generalized Forward-Inverse (GFI) framework, building on the assumption of manifolds and latent space translations. We show that GFI encompasses previous works in deep learning for subsurface imaging, which can be viewed as specific instantiations of GFI. We also propose two new model architectures within the framework of GFI: Latent U-Net and Invertible X-Net, leveraging the power of U-Nets for domain translation and the ability of IU-Nets to simultaneously learn forward and inverse translations, respectively. We show that our proposed models achieve state-of-the-art (SOTA) performance for forward and inverse problems on a wide range of synthetic datasets, and also investigate their zero-shot effectiveness on two real-world-like datasets. The code is available at https://github.com/KGML-lab/Generalized-Forward-Inverse-Framework-for-DL4SI

1895. Interference Among First-Price Pacing Equilibria: A Bias and Variance Analysis

链接：https://iclr.cc/virtual/2025/poster/30875 abstract： A/B testing is widely used in the internet industry. For online marketplaces (such as advertising markets), standard approaches to A/B testing may lead to biased results when buyers have budget constraints, as budget consumption in one arm of the experiment impacts performance of the other arm. This is often addressed using a budget-split design. Yet such splitting may degrade statistical performance as budgets become too small in each arm.We propose a parallel budget-controlled A/B testing design where we use market segmentation to identify submarkets in the larger market, and we run parallel budget-split experiments in each submarket.We demonstrate the effectiveness of this approach on real experiments on advertising markets at Meta.Then, we formally study interference that derives from such experimental designs, using the first-price pacing equilibrium framework as our model of market equilibration.We propose a debiased surrogate that eliminates the first-order bias of FPPE, and derive a plug-in estimator for the surrogate and establish its asymptotic normality. We then provide an estimation procedure for submarket parallel budget-controlled A/B tests. Finally, we present numerical examples on semi-synthetic data, confirming that the debiasing technique achieves the desired coverage properties.

1896. MAGE: Model-Level Graph Neural Networks Explanations via Motif-based Graph Generation

链接：https://iclr.cc/virtual/2025/poster/27883 abstract： Graph Neural Networks (GNNs) have shown remarkable success in molecular tasks, yet their interpretability remains challenging. Traditional model-level explanation methods like XGNN and GNNInterpreter often fail to identify valid substructures like rings, leading to questionable interpretability. This limitation stems from XGNN's atom-by-atom approach and GNNInterpreter's reliance on average graph embeddings, which overlook the essential structural elements crucial for molecules. To address these gaps, we introduce an innovative Motif-bAsed GNN Explainer (MAGE) that uses motifs as fundamental units for generating explanations. Our approach begins with extracting potential motifs through a motif decomposition technique. Then, we utilize an attention-based learning method to identify class-specific motifs. Finally, we employ a motif-based graph generator for each class to create molecular graph explanations based on these class-specific motifs. This novel method not only incorporates critical substructures into the explanations but also guarantees their validity, yielding results that are human-understandable. Our proposed method's effectiveness is demonstrated through quantitative and qualitative assessments conducted on six real-world molecular datasets.

1897. Transformers Struggle to Learn to Search

链接：https://iclr.cc/virtual/2025/poster/30677 abstract： Search is an ability foundational in many important tasks, and recent studies have shown that large language models (LLMs) struggle to perform search robustly. It is unknown whether this inability is due to a lack of data, insufficient model parameters, or fundamental limitations of the transformer architecture. In this work, we use the foundational graph connectivity problem as a testbed to generate effectively limitless high-coverage data to train small transformers and test whether they can learn to perform search. We find that, when given the right training distribution, the transformer is able to learn to search.We analyze the algorithm that the transformer has learned through a novel mechanistic interpretability technique that enables us to extract the computation graph from the trained model. We find that for each vertex in the input graph, transformers compute the set of vertices reachable from that vertex. Each layer then progressively expands these sets, allowing the model to search over a number of vertices exponential in the number of layers.However, we find that as the input graph size increases, the transformer has greater difficulty in learning the task. This difficulty is not resolved even as the number of parameters is increased, suggesting that increasing model scale will not lead to robust search abilities. We also find that performing search in-context (i.e., chain-of-thought) does not resolve this inability to learn to search on larger graphs.

1898. Adversarial Latent Feature Augmentation for Fairness

链接：https://iclr.cc/virtual/2025/poster/29055 abstract： Achieving fairness in machine learning remains a critical challenge, especially due to the opaque effects of data augmentation on input spaces within nonlinear neural networks. Nevertheless, current approaches that emphasize augmenting latent features, rather than input spaces, offer limited insights into their ability to detect and mitigate bias. In response, we introduce the concept of the "unfair region" in the latent space, a subspace that highlights areas where misclassification rates for certain demographic groups are disproportionately high, leading to unfair prediction results. To address this, we propose Adversarial Latent Feature Augmentation (ALFA), a method that leverages adversarial fairness attacks to perturb latent space features, which are then used as data augmentation for fine-tuning. ALFA intentionally shifts latent features into unfair regions, and the last layer of the network is fine-tuned with these perturbed features, leading to a corrected decision boundary that enhances fairness in classification in a cost-effective manner. We present a theoretical framework demonstrating that our adversarial fairness objective reliably generates biased feature perturbations, and that fine-tuning on samples from these unfair regions ensures fairness improvements. Extensive experiments across diverse datasets, modalities, and backbone networks validate that training with these adversarial features significantly enhances fairness while maintaining predictive accuracy in classification tasks.

1899. Simple Guidance Mechanisms for Discrete Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/28719 abstract： Diffusion models for continuous data gained widespread adoption owing to their high quality generation and control mechanisms. However, controllable diffusion on discrete data faces challenges given that continuous guidance methods do not directly apply to discrete diffusion. Here, we provide a straightforward derivation of classifier-free and classifier-based guidance for discrete diffusion, as well as a new class of diffusion models that leverage uniform noise and that are more guidable because they can continuously edit their outputs. We improve the quality of these models with a novel continuous-time variational lower bound that yields state-of-the-art performance, especially in settings involving guidance or fast generation. Empirically, we demonstrate that our guidance mechanisms combined with uniform noise diffusion improve controllable generation relative to autoregressive and diffusion baselines on several discrete data domains, including genomic sequences, small molecule design, and discretized image generation.

1900. Layer Swapping for Zero-Shot Cross-Lingual Transfer in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/27913 abstract： Model merging, such as model souping, is the practice of combining different models with the same architecture together without further training. In this work, we present a model merging methodology that addresses the difficulty of fine-tuning Large Language Models (LLMs) for target tasks in non-English languages, where task-specific data is often unavailable. We focus on mathematical reasoning and without in-language math data, facilitate cross-lingual transfer by composing language and math capabilities. Starting from the same pretrained model, we fine-tune separate "experts" on math instruction data in English and on generic instruction data in the target language. We then replace the top and bottom transformer layers of the math expert directly with layers from the language expert, which consequently enhances math performance in the target language. The resulting merged models outperform the individual experts and other merging methods on the math benchmark, MGSM, by 10% across four major languages where math instruction data is scarce. In addition, this layer swapping is simple, inexpensive, and intuitive, as it is based on an interpretative analysis of the most important parameter changes during the fine-tuning of each expert. The ability to successfully re-compose LLMs for cross-lingual transfer in this manner opens up future possibilities to combine model expertise, create modular solutions, and transfer reasoning capabilities across languages all post hoc.

1901. Improved Finite-Particle Convergence Rates for Stein Variational Gradient Descent

链接：https://iclr.cc/virtual/2025/poster/28115 abstract： We provide finite-particle convergence rates for the Stein Variational

Gradient Descent (SVGD) algorithm in the Kernelized Stein Discrepancy ($\KSD$) and Wasserstein-2 metrics. Our key insight is that the time derivative of the relative entropy between the joint density of N particle locations and the N-fold product target measure, starting from a regular initial distribution, splits into a dominant 'negative part' proportional to N times the expected $\KSD^2$ and a smaller 'positive part'. This observation leads to $\KSD$ rates of order $1/\sqrt{N}$, in both continuous and discrete time, providing a near optimal (in the sense of matching the corresponding i.i.d. rates) double exponential improvement over the recent result by~\cite{shi2024finite}. Under mild assumptions on the kernel and potential, these bounds also grow polynomially in the dimension d. By adding a bilinear component to the kernel, the above approach is used to further obtain Wasserstein-2 convergence in continuous time. For the case of `bilinear + Mat\'ern' kernels, we derive Wasserstein-2 rates that exhibit a curse-of-dimensionality similar to the i.i.d. setting. We also obtain marginal convergence and long-time propagation of chaos results for the time-averaged particle laws.

1902. Learning Randomized Algorithms with Transformers

链接：https://iclr.cc/virtual/2025/poster/29471 abstract： Randomization is a powerful tool that endows algorithms with remarkable properties. For instance, randomized algorithms excel in adversarial settings, often surpassing the worst-case performance of deterministic algorithms with large margins. Furthermore, their success probability can be amplified by simple strategies such as repetition and majority voting. In this paper, we enhance deep neural networks, in particular transformer models, with randomization. We demonstrate for the first time that randomized algorithms can be instilled in transformers through learning, in a purely data- and objective-driven manner. First, we analyze known adversarial objectives for which randomized algorithms offer a distinct advantage over deterministic ones. We then show that common optimization techniques, such as gradient descent or evolutionary strategies, can effectively learn transformer parameters that make use of the randomness provided to the model. To illustrate the broad applicability of randomization in empowering neural networks, we study three conceptual tasks: associative recall, graph coloring, and agents that explore grid worlds. In addition to demonstrating increased robustness against oblivious adversaries through learned randomization, our experiments reveal remarkable performance improvements due to the inherently random nature of the neural networks' computation and predictions.

1903. Encryption-Friendly LLM Architecture

链接：https://iclr.cc/virtual/2025/poster/28288 abstract： Large language models (LLMs) offer personalized responses based on user interactions, but this use case raises serious privacy concerns. Homomorphic encryption (HE) is a cryptographic protocol supporting arithmetic computations in encrypted states and provides a potential solution for privacy-preserving machine learning (PPML). However, the computational intensity of transformers poses challenges for applying HE to LLMs. In this work, we propose a modified HE-friendly transformer architecture with an emphasis on inference following personalized (private) fine-tuning. Utilizing LoRA fine-tuning and Gaussian kernels, we achieve significant computational speedups---6.94$\times$ for fine-tuning and 2.3$\times$ for inference---while maintaining performance comparable to plaintext models. Our findings provide a viable proof of concept for offering privacy-preserving LLM services in areas where data protection is crucial. Our code is available on GitHub.

1904. Bonsai: Gradient-free Graph Condensation for Node Classification

链接：https://iclr.cc/virtual/2025/poster/30924 abstract： Graph condensation has emerged as a promising avenue to enable scalable training of GNNs by compressing the training dataset while preserving essential graph characteristics. Our study uncovers significant shortcomings in current graph condensation techniques. First, the majority of the algorithms paradoxically require training on the full dataset to perform condensation. Second, due to their gradient-emulating approach, these methods require fresh condensation for any change in hyperparameters or GNN architecture, limiting their flexibility and reusability. To address these challenges, we present Bonsai, a novel graph condensation method empowered by the observation that *computation trees* form the fundamental processing units of message-passing GNNs. Bonsai condenses datasets by encoding a careful selection of *exemplar* trees that maximize the representation of all computation trees in the training set. This unique approach imparts Bonsai as the first linear-time, model-agnostic graph condensation algorithm for node classification that outperforms existing baselines across 7 real-world datasets on accuracy, while being 22 times faster on average. Bonsai is grounded in rigorous mathematical guarantees on the adopted approximation strategies, making it robust to GNN architectures, datasets, and parameters.

1905. Sensitivity Verification for Additive Decision Tree Ensembles

链接：https://iclr.cc/virtual/2025/poster/28785 abstract： Tree ensemble models, such as Gradient Boosted Decision Trees (GBDTs) and random forests, are widely popular models for a variety of machine learning tasks. The power of these models comes from the ensemble of decision trees, which makes analysis of such models significantly harder than for single trees. As a result, recent work has focused on developing exact and approximate techniques for questions such as robustness verification, fairness and explainability for such models of tree ensembles.In this paper, we focus on a specific problem of feature sensitivity for additive decision tree ensembles and build a formal verification framework for a parametrized variant of it, where we also take into account the confidence of the tree ensemble in its output. We start by showing theoretical (NP-)hardness of the problem and explain how it relates to other verification problems. Next, we provide a novel encoding of the problem using pseudo-Boolean constraints. Based on this encoding, we develop a tunable algorithm to perform sensitivity analysis, which can trade off precision for running time. We implement our algorithm and study its performance on a suite of GBDT benchmarks from the literature. Our experiments show the practical utility of our approach and its improved performance compared to existing

approaches.

1906. Can LLMs Really Learn to Translate a Low-Resource Language from One Grammar Book?

链接：https://iclr.cc/virtual/2025/poster/29173 abstract： Extremely low-resource (XLR) languages lack substantial corpora for training NLP models, motivating the use of all available resources such as dictionaries and grammar books. Machine Translation from One Book (Tanzer et al., 2024) suggests that prompting long-context LLMs with one grammar book enables English–Kalamang translation, an XLR language unseen by LLMs—a noteworthy case of linguistics helping an NLP task. We investigate the source of this translation ability, finding almost all improvements stem from the book's parallel examples rather than its grammatical explanations. We find similar results for Nepali and Guarani, seen low-resource languages, and we achieve performance comparable to an LLM with a grammar book by simply fine-tuning an encoder-decoder translation model. We then investigate where grammar books help by testing two linguistic tasks, grammaticality judgment and gloss prediction, and we explore what kind of grammatical knowledge helps by introducing a typological feature prompt that achieves leading results on these more relevant tasks. We thus emphasise the importance of task-appropriate data for XLR languages: parallel examples for translation, and grammatical data for linguistic tasks. As we find no evidence that long-context LLMs can make effective use of grammatical explanations for XLR translation, we conclude data collection for multilingual XLR tasks such as translation is best focused on parallel data over linguistic description.

1907. An Undetectable Watermark for Generative Image Models

链接：https://iclr.cc/virtual/2025/poster/28619 abstract： We present the first undetectable watermarking scheme for generative image models.Undetectability ensures that no efficient adversary can distinguish between watermarked and un-watermarked images, even after making many adaptive queries.In particular, an undetectable watermark does not degrade image quality under any efficiently computable metric.Our scheme works by selecting the initial latents of a diffusion model using a pseudorandom error-correcting code (Christ and Gunn, 2024), a strategy which guarantees undetectability and robustness. We experimentally demonstrate that our watermarks are quality-preserving and robust using Stable Diffusion 2.1.Our experiments verify that, in contrast to every prior scheme we tested, our watermark does not degrade image quality.Our experiments also demonstrate robustness: existing watermark removal attacks fail to remove our watermark from images without significantly degrading the quality of the images.Finally, we find that we can robustly encode 512 bits in our watermark, and up to 2500 bits when the images are not subjected to watermark removal attacks.Our code is available at https://github.com/XuandongZhao/PRC-Watermark.

1908. Diffusion Models and Gaussian Flow Matching: Two Sides of the Same Coin

链接：https://iclr.cc/virtual/2025/poster/31359 abstract： Flow matching and diffusion models are two popular frameworks in generative modeling. Despite seeming similar, there is some confusion in the community about their exact connection. In this post we aim to clear up this confusion and show that diffusion models and Gaussian flow matching are the same -- Different model specifications lead to different noise schedules and loss weightings but correspond to the same generative model. That's great news, it means that you can use the two frameworks interchangeably.

1909. Differentiable Causal Discovery for Latent Hierarchical Causal Models

链接：https://iclr.cc/virtual/2025/poster/30546 abstract： Discovering causal structures with latent variables from observational data is a fundamental challenge in causal discovery. Existing methods often rely on constraint-based, iterative discrete searches, limiting their scalability for large numbers of variables. Moreover, these methods frequently assume linearity or invertibility, restricting their applicability to real-world scenarios. We present new theoretical results on the identifiability of non-linear latent hierarchical causal models, relaxing previous assumptions in the literature about the deterministic nature of latent variables and exogenous noise. Building on these insights, we develop a novel differentiable causal discovery algorithm that efficiently estimates the structure of such models. To the best of our knowledge, this is the first work to propose a differentiable causal discovery method for non-linear latent hierarchical models. Our approach outperforms existing methods in both accuracy and scalability. Furthermore, we demonstrate its practical utility by learning interpretable hierarchical latent structures from high-dimensional image data and demonstrate its effectiveness on downstream tasks such as transfer learning.

1910. On the Relation between Trainability and Dequantization of Variational Quantum Learning Models

链接：https://iclr.cc/virtual/2025/poster/29529 abstract： Quantum machine learning (QML) explores the potential advantages of quantum computers for machine learning tasks, with variational QML among the main current approaches.While quantum computers promise to solve problems that are classically intractable, it has been recently shown that a particular quantum algorithm which outperforms all pre-existing classical algorithms can be matched by a newly developed classical approach (often inspired by the quantum algorithm).We say such algorithms have been dequantized.For QML models to be effective, they must be trainable and non-dequantizable.The relationship between these properties is still not fully understood and recent works

raised into question to what extent we could ever have QML models which are both trainable and non-dequantizable.This challenges the potential of QML altogether.In this work we answer open questions regarding when trainability and non-dequantization are compatible.We first formalize the key concepts and put them in the context of prior research.We introduce the role of "variationalness" of QML models using well-known quantum circuit architectures as leading examples.Our results provide recipes for variational QML models that are trainable and non-dequantizable.By ensuring that variational QML models are both trainable and non-dequantizable, we pave the way toward practical relevance.

1911. Sketching for Convex and Nonconvex Regularized Least Squares with Sharp Guarantees

链接：https://iclr.cc/virtual/2025/poster/30805 abstract： Randomized algorithms play a crucial role in efficiently solving large-scale optimization problems. In this paper, we introduce Sketching for Regularized Optimization (SRO), a fast sketching algorithm designed for least squares problems with convex or nonconvex regularization. SRO operates by first creating a sketch of the original data matrix and then solving the sketched problem. We establish minimax optimal rates for sparse signal estimation by addressing the sketched sparse convex and nonconvex learning problems. Furthermore, we propose a novel Iterative SRO algorithm, which reduces the approximation error geometrically for sketched convex regularized problems. To the best of our knowledge, this work is among the first to provide a unified theoretical framework demonstrating minimax rates for convex and nonconvex sparse learning problems via sketching. Experimental results validate the efficiency and effectiveness of both the SRO and Iterative SRO algorithms.

1912. Lawma: The Power of Specialization for Legal Annotation

链接：https://iclr.cc/virtual/2025/poster/30835 abstract： Annotation and classification of legal text are central components of empirical legal research. Traditionally, these tasks are often delegated to trained research assistants. Motivated by the advances in language modeling, empirical legal scholars are increasingly turning to commercial models, hoping that it will alleviate the significant cost of human annotation. In this work, we present a comprehensive analysis of large language models' current abilities to perform legal annotation tasks. To do so, we construct CaselawQA, a benchmark comprising 260 legal text classification tasks, nearly all new to the machine learning community. We demonstrate that commercial models, such as GPT-4.5 and Claude 3.7 Sonnet, achieve non-trivial accuracy but generally fall short of the performance required for legal work. We then demonstrate that small, lightly fine-tuned models vastly outperform commercial models. A few dozen to a few hundred labeled examples are usually enough to achieve higher accuracy. Our work points to a viable alternative to the predominant practice of prompting commercial models. For concrete legal annotation tasks with some available labeled data, researchers are likely better off using a fine-tuned open-source model. Code, datasets, and fine-tuned models are available at https://github.com/socialfoundations/lawma.

1913. Fengbo: a Clifford Neural Operator pipeline for 3D PDEs in Computational Fluid Dynamics

链接：https://iclr.cc/virtual/2025/poster/29396 abstract： We introduce Fengbo, a pipeline entirely in Clifford Algebra to solve 3D partial differential equations (PDEs) specifically for computational fluid dynamics (CFD). Fengbo is an architecture composed of only 3D convolutional and Fourier Neural Operator (FNO) layers, all working in 3D Clifford Algebra. It models the PDE solution problem as an interpretable mapping from the geometry to the physics of the problem. Despite having just few layers, Fengbo achieves competitive accuracy, superior to 5 out of 6 proposed models reported in \cite{li2024geometry} for the $\emph{ShapeNet Car}$ dataset, and it does so with only 42 million trainable parameters, at a reduced computational complexity compared to graph-based methods, and estimating jointly pressure \emph{and} velocity fields. In addition, the output of each layer in Fengbo can be clearly visualised as objects and physical quantities in 3D space, making it a whitebox model. By leveraging Clifford Algebra and establishing a direct mapping from the geometry to the physics of the PDEs, Fengbo provides an efficient, geometry- and physics-aware approach to solving complex PDEs.

1914. Problem-Parameter-Free Federated Learning

链接：https://iclr.cc/virtual/2025/poster/29194 abstract： Federated learning (FL) has garnered significant attention from academia and industry in recent years due to its advantages in data privacy, scalability, and communication efficiency. However, current FL algorithms face a critical limitation: their performance heavily depends on meticulously tuned hyperparameters, particularly the learning rate or stepsize. This manual tuning process is challenging in federated settings due to data heterogeneity and limited accessibility of local datasets. Consequently, the reliance on problem-specific parameters hinders the widespread adoption of FL and potentially compromises its performance in dynamic or diverse environments. To address this issue, we introduce PAdaMFed, a novel algorithm for nonconvex FL that carefully combines adaptive stepsize and momentum techniques. PAdaMFed offers two key advantages: 1) it operates autonomously without relying on problem-specific parameters; and 2) it manages data heterogeneity and partial participation without requiring heterogeneity bounds. Despite these benefits, PAdaMFed provides several strong theoretical guarantees: 1) It achieves state-of-the-art convergence rates with a sample complexity of $\mathcal{O}(\epsilon^{-4})$ and communication complexity of $\mathcal{O}(\epsilon^{-3})$ to obtain an accuracy of $\|\nabla f\left(\boldsymbol{\theta}\right)\| \leq \epsilon$, even using constant learning rates; 2) these complexities can be improved to the best-known $\mathcal{O}(\epsilon^{-3})$ for sampling and $\mathcal{O}(\epsilon^{-2})$ for communication when incorporating variance reduction; 3) it exhibits linear speedup with respect to the number of local update steps and participating

clients at each global round. These attributes make PAdaMFed highly scalable and adaptable for various real-world FL applications. Extensive empirical evidence on both image classification and sentiment analysis tasks validates the efficacy of our approaches.

1915. Efficiently Learning at Test-Time: Active Fine-Tuning of LLMs

链接：https://iclr.cc/virtual/2025/poster/29875 abstract： Recent efforts in fine-tuning language models often rely on automatic data selection, commonly using Nearest Neighbors retrieval from large datasets.However, we theoretically show that this approach tends to select redundant data, limiting its effectiveness or even hurting performance.To address this, we introduce SIFT, a data selection algorithm designed to reduce uncertainty about the model's response given a prompt, which unifies ideas from retrieval and active learning.Whereas Nearest Neighbor retrieval typically fails in the presence of information duplication, SIFT accounts for information duplication and optimizes the overall information gain of the selected examples.We focus our evaluations on fine-tuning at test-time for prompt-specific language modeling on the Pile dataset, and show that SIFT consistently outperforms Nearest Neighbor retrieval, with minimal computational overhead.Moreover, we show that our uncertainty estimates can predict the performance gain of test-time fine-tuning, and use this to develop an adaptive algorithm that invests test-time compute proportional to realized performance gains.We provide the activeft (Active Fine-Tuning) library which can be used as a drop-in replacement for Nearest Neighbor retrieval.

1916. LoRA-X: Bridging Foundation Models with Training-Free Cross-Model Adaptation

链接：https://iclr.cc/virtual/2025/poster/30873 abstract： The rising popularity of large foundation models has led to a heightened demand for parameter-efficient fine-tuning methods, such as Low-Rank Adaptation (LoRA), which offer performance comparable to full model fine-tuning while requiring only a few additional parameters tailored to the specific base model. When such base models are deprecated and replaced, all associated LoRA modules must be retrained, requiring access to either the original training data or a substantial amount of synthetic data that mirrors the original distribution. However, the original data is often inaccessible due to privacy or licensing issues, and generating synthetic data may be impractical and insufficiently representative. These factors complicate the fine-tuning process considerably. To address this challenge, we introduce a new adapter, Cross-Model Low-Rank Adaptation (LoRA-X), which enables the training-free transfer of LoRA parameters across source and target models, eliminating the need for original or synthetic training data. Our approach imposes the adapter to operate within the subspace of the source base model. This constraint is necessary because our prior knowledge of the target model is limited to its weights, and the criteria for ensuring the adapter’s transferability are restricted to the target base model’s weights and subspace. To facilitate the transfer of LoRA parameters of the source model to a target model, we employ the adapter only in the layers of the target model that exhibit an acceptable level of subspace similarity. Our extensive experiments demonstrate the effectiveness of LoRA-X for text-to-image generation, including Stable Diffusion v1.5 and Stable Diffusion XL.

1917. Beyond FVD: An Enhanced Evaluation Metrics for Video Generation Distribution Quality

链接：https://iclr.cc/virtual/2025/poster/29066 abstract： The Fréchet Video Distance (FVD) is a widely adopted metric for evaluating video generation distribution quality. However, its effectiveness relies on critical assumptions. Our analysis reveals three significant limitations: (1) the non-Gaussianity of the Inflated 3D Convnet (I3D) feature space; (2) the insensitivity of I3D features to temporal distortions; (3) the impractical sample sizes required for reliable estimation. These findings undermine FVD's reliability and show that FVD falls short as a standalone metric for video generation evaluation. After extensive analysis of a wide range of metrics and backbone architectures, we propose JEDi, the JEPA Embedding Distance, based on features derived from a Joint Embedding Predictive Architecture, measured using Maximum Mean Discrepancy with polynomial kernel. Our experiments on multiple open-source datasets show clear evidence that it is a superior alternative to the widely used FVD metric, requiring only 16% of the samples to reach its steady value, while increasing alignment with human evaluation by 34%, on average.Project page: https://oooolga.github.io/JEDi.github.io/.

1918. Dual Process Learning: Controlling Use of In-Context vs. In-Weights Strategies with Weight Forgetting

链接：https://iclr.cc/virtual/2025/poster/28644 abstract： Language models have the ability to perform in-context learning (ICL), allowing them to flexibly adapt their behavior based on context. This contrasts with in-weights learning (IWL), where memorized information is encoded in model parameters after iterated observations of data. An ideal model should be able to flexibly deploy both of these abilities. Despite their apparent ability to learn in-context, language models are known to struggle when faced with unseen or rarely seen tokens (Land & Bartolo, 2024). Hence, we study $\textbf{structural in-context learning}$, which we define as the ability of a model to execute in-context learning on arbitrary novel tokens -- so called because the model must generalize on the basis of e.g. sentence structure or task structure, rather than content encoded in token embeddings. We study structural in-context algorithms on both synthetic and naturalistic tasks using toy models, masked language models, and autoregressive language models. We find that structural ICL appears before quickly disappearing early in LM pretraining. While it has been shown that ICL can diminish during training (Singh et al., 2023), we find that prior work does not account for structural ICL. Building on Chen et al. (2024) 's active forgetting method, we introduce pretraining and finetuning methods that can modulate the

preference for structural ICL and IWL. Importantly, this allows us to induce a $\textit{dual process strategy}$ where in-context and in-weights solutions coexist within a single model.

1919. VOILA: Evaluation of MLLMs For Perceptual Understanding and Analogical Reasoning

链接：https://iclr.cc/virtual/2025/poster/28264 abstract： Multimodal Large Language Models (MLLMs) have become a powerful tool for integrating visual and textual information. Despite their exceptional performance on visual understanding benchmarks, measuring their ability to reason abstractly across multiple images remains a significant challenge. To address this, we introduce VOILA, a large-scale, open-ended, dynamic benchmark designed to evaluate MLLMs' perceptual understanding and abstract relational reasoning. VOILA employs an analogical mapping approach in the visual domain, requiring models to generate an image that completes an analogy between two given image pairs, reference and application, without relying on predefined choices. Our experiments demonstrate that the analogical reasoning tasks in VOILA present a challenge to MLLMs. Through multi-step analysis, we reveal that current MLLMs struggle to comprehend inter-image relationships and exhibit limited capabilities in high-level relational reasoning. Notably, we observe that performance improves when following a multi-step strategy of least-to-most prompting. Comprehensive evaluations on open-source models and GPT-4o show that on text-based answers, the best accuracy for challenging scenarios is 13% (LLaMa 3.2) and even for simpler tasks is only 29% (GPT-4o), while human performance is significantly higher at 70% across both difficulty levels.

1920. Differentiable Optimization of Similarity Scores Between Models and Brains

链接：https://iclr.cc/virtual/2025/poster/27905 abstract： How do we know if two systems - biological or artificial - process information in a similar way? Similarity measures such as linear regression, Centered Kernel Alignment (CKA), Normalized Bures Similarity (NBS), and angular Procrustes distance, are often used to quantify this similarity. However, it is currently unclear what drives high similarity scores and even what constitutes a "good" score. Here, we introduce a novel tool to investigate these questions by differentiating through similarity measures to directly maximize the score. Surprisingly, we find that high similarity scores do not guarantee encoding task-relevant information in a manner consistent with neural data; and this is particularly acute for CKA and even some variations of cross-validated and regularized linear regression. We find no consistent threshold for a good similarity score - it depends on both the measure and the dataset. In addition, synthetic datasets optimized to maximize similarity scores initially learn the highest variance principal component of the target dataset, but some methods like angular Procrustes capture lower variance dimensions much earlier than methods like CKA. To shed light on this, we mathematically derive the sensitivity of CKA, angular Procrustes, and NBS to the variance of principal component dimensions, and explain the emphasis CKA places on high variance components. Finally, by jointly optimizing multiple similarity measures, we characterize their allowable ranges and reveal that some similarity measures are more constraining than others. While current measures offer a seemingly straightforward way to quantify the similarity between neural systems, our work underscores the need for careful interpretation. We hope the tools we developed will be used by practitioners to better understand current and future similarity measures.

1921. Progressive distillation induces an implicit curriculum

链接：https://iclr.cc/virtual/2025/poster/27848 abstract： Knowledge distillation leverages a teacher model to improve the training of a student model. A persistent challenge is that a better teacher does not always yield a better student, to which a common mitigation is to use additional supervision from several “intermediate” teachers. One empirically validated variant of this principle is progressive distillation, where the student learns from successive intermediate checkpoints of the teacher. Using sparse parity as a sandbox, we identify an implicit curriculum as one mechanism through which progressive distillation accelerates the student’s learning. This curriculum is available only through the intermediate checkpoints but not the final converged one, and imparts both empirical acceleration and a provable sample complexity benefit to the student. We then extend our investigation to Transformers trained on probabilistic context-free grammars (PCFGs) and real-world pre-training datasets (Wikipedia and Books). Through probing the teacher model, we identify an analogous implicit curriculum where the model progressively learns features that capture longer context. Our theoretical and empirical findings on sparse parity, complemented by empirical observations on more complex tasks, highlight the benefit of progressive distillation via implicit curriculum across setups.

1922. Enhanced Diffusion Sampling via Extrapolation with Multiple ODE Solutions

链接：https://iclr.cc/virtual/2025/poster/28214 abstract： Diffusion probabilistic models (DPMs), while effective in generating high-quality samples, often suffer from high computational costs due to their iterative sampling process. To address this, we propose an enhanced ODE-based sampling method for DPMs inspired by Richardson extrapolation, which reduces numerical error and improves convergence rates. Our method, RX-DPM, leverages multiple ODE solutions at intermediate time steps to extrapolate the denoised prediction in DPMs. Our method, RX-DPM, leverages multiple ODE solutions at intermediate time steps to extrapolate the denoised prediction in DPMs. This significantly enhances the accuracy of estimations for the final sample while maintaining the number of function evaluations (NFEs). Unlike standard Richardson extrapolation, which assumes

uniform discretization of the time grid, we develop a more general formulation tailored to arbitrary time step scheduling, guided by local truncation error derived from a baseline sampling method. The simplicity of our approach facilitates accurate estimation of numerical solutions without significant computational overhead, and allows for seamless and convenient integration into various DPMs and solvers. Additionally, RX-DPM provides explicit error estimates, effectively demonstrating the faster convergence as the leading error term's order increases. Through a series of experiments, we show that the proposed method improves the quality of generated samples without requiring additional sampling iterations.

1923. McEval: Massively Multilingual Code Evaluation

链接：https://iclr.cc/virtual/2025/poster/29445 abstract： Code large language models (LLMs) have shown remarkable advances in code understanding, completion, and generation tasks. Programming benchmarks, comprised of a selection of code challenges and corresponding test cases, serve as a standard to evaluate the capability of different LLMs in such tasks. However, most existing benchmarks primarily focus on Python and are still restricted to a limited number of languages, where other languages are translated from the Python samples degrading the data diversity. To further facilitate the research of code LLMs, we propose a massively multilingual code benchmark covering 40 programming languages (McEval) with 16K test samples, which substantially pushes the limits of code LLMs in multilingual scenarios. The benchmark contains challenging code completion, understanding, and generation evaluation tasks with finely curated massively multilingual instruction corpora McEval-Instruct. In addition, we introduce an effective multilingual coder mCoder trained on McEval-Instruct to support multilingual programming language generation. Extensive experimental results on McEval show that there is still a difficult journey between open-source models and closed-source LLMs in numerous languages. The instruction corpora and evaluation benchmark are available at https://github.com/MCEVAL/McEval.

1924. E(n) Equivariant Topological Neural Networks

链接：https://iclr.cc/virtual/2025/poster/30602 abstract： Graph neural networks excel at modeling pairwise interactions, but they cannot flexibly accommodate higher-order interactions and features. Topological deep learning (TDL) has emerged recently as a promising tool for addressing this issue. TDL enables the principled modeling of arbitrary multi-way, hierarchical higher-order interactions by operating on combinatorial topological spaces, such as simplicial or cell complexes, instead of graphs. However, little is known about how to leverage geometric features such as positions and velocities for TDL. This paper introduces E(n)-Equivariant Topological Neural Networks (ETNNs), which are E(n)-equivariant message-passing networks operating on combinatorial complexes, formal objects unifying graphs, hypergraphs, simplicial, path, and cell complexes. ETNNs incorporate geometric node features while respecting rotation, reflection, and translation equivariance. Moreover, being TDL models, ETNNs are natively ready for settings with heterogeneous interactions. We provide a theoretical analysis to show the improved expressiveness of ETNNs over architectures for geometric graphs. We also show how E(n)-equivariant variants of TDL models can be directly derived from our framework. The broad applicability of ETNNs is demonstrated through two tasks of vastly different scales: i) molecular property prediction on the QM9 benchmark and ii) land-use regression for hyper-local estimation of air pollution with multi-resolution irregular geospatial data. The results indicate that ETNNs are an effective tool for learning from diverse types of richly structured data, as they match or surpass SotA equivariant TDL models with a significantly smaller computational burden, thus highlighting the benefits of a principled geometric inductive bias. Our implementation of ETNNs can be found at https://github.com/NSAPH-Projects/topological-equivariant-networks.

1925. Tell me about yourself: LLMs are aware of their learned behaviors

链接：https://iclr.cc/virtual/2025/poster/30155 abstract：

1926. Conformal Language Model Reasoning with Coherent Factuality

链接：https://iclr.cc/virtual/2025/poster/30640 abstract： Language models are increasingly being used in important decision pipelines, so ensuring the correctness of their outputs is crucial. Recent work has proposed evaluating the “factuality” of claims decomposed from a language model generation and applying conformal prediction techniques to filter out those claims that are not factual. This can be effective for tasks such as information retrieval, where constituent claims may be evaluated in isolation for factuality, but is not appropriate for reasoning tasks, as steps of a logical argument can be evaluated for correctness only within the context of the claims that precede them. To capture this, we define “coherent factuality” and develop a conformal-prediction-based method to guarantee coherent factuality for language model outputs. Our approach applies split conformal prediction to subgraphs within a "deducibility" graph that represents the steps of a reasoning problem. We evaluate our method on mathematical reasoning problems from the MATH and FELM datasets and find that our algorithm consistently produces correct and substantiated orderings of claims, achieving coherent factuality across target coverage levels. Moreover, we achieve 90\% factuality on our stricter definition while retaining 80\% or more of the original claims, highlighting the utility of our deducibility-graph-guided approach.

1927. Better than Your Teacher: LLM Agents that learn from Privileged AI Feedback

链接：https://iclr.cc/virtual/2025/poster/28093 abstract： While large language models (LLMs) show impressive decision-making abilities, current methods lack a mechanism for automatic self-improvement from errors during task execution. We

propose LEAP, an iterative fine-tuning framework that continually improves LLM agents using feedback from AI expert teachers. Our key insight is to equip the expert teachers with a privileged state -- information available during training but hidden at test time. This allows even weak experts to provide precise guidance, significantly improving the student agent's performance without access to privileged information at test time.We evaluate LEAP on multiple decision-making benchmarks, including text-based games (ALFWorld), web navigation (WebShop), and interactive coding (Intercode Bash). Our experiments show that LEAP (1) outperforms behavior cloning and ReAct baselines (2) enables weak student models (e.g., Llama3-8B) to exceed performance of strong teacher models (GPT-4o), and (3) allows weak models to self-improve using privileged versions of themselves. We provide a theoretical analysis showing that LEAP's success hinges on balancing privileged information with student's realizability, which we empirically validate. Our code is available at \url{https://leap-llm.github.io}.

1928. Finding and Only Finding Differential Nash Equilibria by Both Pretending to be a Follower

链接：https://iclr.cc/virtual/2025/poster/31513 abstract： Finding Nash equilibria in two-player differentiable games is a classical problem in game theory with important relevance in machine learning. We propose double Follow-the-Ridge (double-FTR), an algorithm that locally converges to and only to differential Nash equilibria in general-sum two-player differentiable games. To our knowledge, double-FTR is the first algorithm with such guarantees for general-sum games. Furthermore, we show that by varying its preconditioner, double-FTR leads to a broader family of algorithms with the same convergence guarantee. In addition, double-FTR avoids oscillation near equilibria due to the real-eigenvalues of its Jacobian at fixed points. Empirically, we validate the double-FTR algorithm on a range of simple zero-sum and general sum games, as well as simple Generative Adversarial Network (GAN) tasks.

1929. TimeSuite: Improving MLLMs for Long Video Understanding via Grounded Tuning

链接：https://iclr.cc/virtual/2025/poster/28421 abstract： Multimodal Large Language Models (MLLMs) have demonstrated impressive performance in short video understanding. However, understanding long-form videos still remains challenging for MLLMs. This paper proposes TimeSuite, a collection of new designs to adapt the existing short-form video MLLMs for long video understanding, including a simple yet efficient framework to process long video sequence, a high-quality video dataset for grounded tuning of MLLMs, and a carefully-designed instruction tuning task to explicitly incorporate the grounding supervision in the traditional QA format. Specifically, based on VideoChat, we propose our long-video MLLM, coined as VideoChat-T, by implementing a token shuffling to compress long video tokens and introducing Temporal Adaptive Position Encoding (TAPE) to enhance the temporal awareness of visual representation. Meanwhile, we introduce the TimePro, a comprehensive grounding-centric instruction tuning dataset composed of 9 tasks and 349k high-quality grounded annotations. Notably, we design a new instruction tuning task type, called Temporal Grounded Caption, to peform detailed video descriptions with the corresponding time stamps prediction. This explicit temporal location prediction will guide MLLM to correctly attend on the visual content when generating description, and thus reduce the hallucination risk caused by the LLMs. Experimental results demonstrate that our TimeSuite provides a successful solution to enhance the long video understanding capability of short-form MLLM, achieving improvement of 5.6% and 6.8% on the benchmarks of Egoschema and VideoMME, respectively. In addition, VideoChat-T exhibits robust zero-shot temporal grounding capabilities, significantly outperforming the existing state-of-the-art MLLMs. After fine-tuning, it performs on par with the traditional supervised expert models.

1930. CarbonSense: A Multimodal Dataset and Baseline for Carbon Flux Modelling

链接：https://iclr.cc/virtual/2025/poster/28534 abstract： Terrestrial carbon fluxes provide vital information about our biosphere's health and its capacity to absorb anthropogenic CO$_2$ emissions. The importance of predicting carbon fluxes has led to the emerging field of data-driven carbon flux modelling (DDCFM), which uses statistical techniques to predict carbon fluxes from biophysical data. However, the field lacks a standardized dataset to promote comparisons between models. To address this gap, we present CarbonSense, the first machine learning-ready dataset for DDCFM. CarbonSense integrates measured carbon fluxes, meteorological predictors, and satellite imagery from 385 locations across the globe, offering comprehensive coverage and facilitating robust model training. Additionally, we provide a baseline model using a current state-of-the-art DDCFM approach and a novel transformer based model. Our experiments illustrate the potential gains that multimodal deep learning techniques can bring to this domain. By providing these resources, we aim to lower the barrier to entry for other deep learning researchers to develop new models and drive new advances in carbon flux modelling.

1931. On the Inherent Privacy Properties of Discrete Denoising Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/31493 abstract： Privacy concerns have led to a surge in the creation of synthetic datasets, with diffusion models emerging as a promising avenue. Although prior studies have performed empirical evaluations on these models, there has been a gap in providing a mathematical characterization of their privacy-preserving capabilities. To address this, we present the pioneering theoretical exploration of the privacy preservation inherent in \emph{discrete diffusion models} (DDMs) for discrete dataset generation. Focusing on per-instance differential privacy (pDP), our framework elucidates

the potential privacy leakage for each data point in a given training dataset, offering insights into how the privacy loss of each point correlates with the dataset's distribution. Our bounds also show that training with s-sized data points leads to a surge in privacy leakage from $(\epsilon, \mathcal{O}(\frac{1}{s^2\epsilon}))$-pDP to $(\epsilon, \mathcal{O}(\frac{1}{s\epsilon}))$-pDP of the DDM during the transition from the pure noise to the synthetic clean data phase, and a faster decay in diffusion coefficients amplifies the privacy guarantee. Finally, we empirically verify our theoretical findings on both synthetic and real-world datasets.

1932. Towards Understanding Why Label Smoothing Degrades Selective Classification and How to Fix It

链接：https://iclr.cc/virtual/2025/poster/30864 abstract： Label smoothing (LS) is a popular regularisation method for training neural networks as it is effective in improving test accuracy and is simple to implement. "Hard" one-hot labels are "smoothed" by uniformly distributing probability mass to other classes, reducing overfitting. Prior work has shown that in some cases LS can degrade selective classification (SC) -- where the aim is to reject misclassifications using a model's uncertainty. In this work, we first demonstrate empirically across an extended range of large-scale tasks and architectures that LS consistently degrades SC. We then address a gap in existing knowledge, providing an explanation for this behaviour by analysing logit-level gradients: LS degrades the uncertainty rank ordering of correct vs incorrect predictions by regularising the max logit more when a prediction is likely to be correct, and less when it is likely to be wrong. This elucidates previously reported experimental results where strong classifiers underperform in SC. We then demonstrate the empirical effectiveness of post-hoc logit normalisation for recovering lost SC performance caused by LS. Furthermore, linking back to our gradient analysis, we again provide an explanation for why such normalisation is effective.

1933. Self-Improving Robust Preference Optimization

链接：https://iclr.cc/virtual/2025/poster/29219 abstract： Online and offline $\mathtt{RLHF}$ methods, such as $\mathtt{PPO}$ and $\mathtt{DPO}$, have been highly successful in aligning AI with human preferences. Despite their success, however, these methods suffer from fundamental limitations: $\mathbf{(a)}$ Models trained with $\mathtt{RLHF}$ can learn from mistakes or negative examples through RL mechanism or contrastive loss during training. However, at inference time, they lack an innate self-improvement mechanism for error corrections. $\mathbf{(b)}$ The optimal solution of existing methods is highly task-dependent, making it difficult for them to generalize to new tasks. To address these challenges, we propose Self-Improving Robust Preference Optimization ($\mathtt{SRPO}$), a practical and mathematically principled offline $\mathtt{RLHF}$ framework. The key idea behind $\mathtt{SRPO}$ is to cast the problem of learning from human preferences as a self-improvement process, mathematically formulated as a min-max objective that jointly optimizes a self-improvement policy and a generative policy in an adversarial fashion. Crucially, the solution for this optimization problem is independent of the training task, which makes it robust to its changes. We then show that this objective can be reformulated as a non-adversarial offline loss, which can be efficiently optimized using standard supervised learning techniques at scale. To demonstrate $\mathtt{SRPO}$'s effectiveness, we evaluate it using AI Win-Rate (WR) against human (GOLD) completions. When tested on the XSum dataset, $\mathtt{SRPO}$ outperforms $\mathtt{DPO}$ by a margin of $\mathbf{15}$% after 5 self-revisions, achieving an impressive $\mathbf{90}$% WR. Moreover, on the challenging Arena-Hard prompts, $\mathtt{SRPO}$ outperforms both $\mathtt{DPO}$ and $\mathtt{IPO}$ (by $\mathbf{4}$% without revision and $\mathbf{6}$% after a single revision), reaching a $\mathbf{56}$% WR against against $\mathtt{Llama-3.1-8B-Instruct}$.

1934. NatureLM-audio: an Audio-Language Foundation Model for Bioacoustics

链接：https://iclr.cc/virtual/2025/poster/28770 abstract： Large language models (LLMs) prompted with text and audio have achieved state-of-the-art performance across various auditory tasks, including speech, music, and general audio, showing emergent abilities on unseen tasks. However, their potential has yet to be fully demonstrated in bioacoustics tasks, such as detecting animal vocalizations in large recordings, classifying rare and endangered species, and labeling context and behavior —tasks that are crucial for conservation, biodiversity monitoring, and animal behavior studies. In this work, we present NatureLM-audio, the first audio-language foundation model specifically designed for bioacoustics. Our training dataset consists of carefully curated text-audio pairs spanning bioacoustics, speech, and music, designed to address the field's limited availability of annotated data. We demonstrate successful transfer of learned representations from music and speech to bioacoustics, and our model shows promising generalization to unseen taxa and tasks. We evaluate NatureLM-audio on a novel benchmark (BEANS-Zero) and it sets a new state of the art on several bioacoustics tasks, including zero-shot classification of unseen species. To advance bioacoustics research, we release our model weights, benchmark data, and open-source the code for training and benchmark data generation and model training.

1935. Synthio: Augmenting Small-Scale Audio Classification Datasets with Synthetic Data

链接：https://iclr.cc/virtual/2025/poster/29111 abstract： We present Synthio, a novel approach for augmenting small-scale audio classification datasets with synthetic data. Our goal is to improve audio classification accuracy with limited labeled data. Traditional data augmentation techniques, which apply artificial transformations (e.g., adding random noise or masking segments), struggle to create data that captures the true diversity present in real-world audios. To address this shortcoming, we

propose to augment the dataset with synthetic audio generated from text-to-audio (T2A) diffusion models. However, synthesizing effective augmentations is challenging because not only should the generated data be acoustically consistent with the underlying small-scale dataset, but they should also have sufficient compositional diversity. To overcome the first challenge, we align the generations of the T2A model with the small-scale dataset using preference optimization. This ensures that the acoustic characteristics of the generated data remain consistent with the small-scale dataset. To address the second challenge, we propose a novel caption generation technique that leverages the reasoning capabilities of Large Language Models to (1) generate diverse and meaningful audio captions and (2) iteratively refine their quality. The generated captions are then used to prompt the aligned T2A model. We extensively evaluate Synthio on ten datasets and four simulated limited-data settings. Results indicate our method consistently outperforms all baselines by 0.1%-39% using a T2A model trained only on weakly-captioned AudioSet.

1936. Non-Adversarial Inverse Reinforcement Learning via Successor Feature Matching

链接：https://iclr.cc/virtual/2025/poster/29965 abstract： In inverse reinforcement learning (IRL), an agent seeks to replicate expert demonstrations through interactions with the environment.Traditionally, IRL is treated as an adversarial game, where an adversary searches over reward models, and a learner optimizes the reward through repeated RL procedures.This game-solving approach is both computationally expensive and difficult to stabilize.In this work, we propose a novel approach to IRL by direct policy search: by exploiting a linear factorization of the return as the inner product of successor features and a reward vector, we design an IRL algorithm by policy gradient descent on the gap between the learner and expert features.Our non-adversarial method does not require learning an explicit reward function and can be solved seamlessly with existing RL algorithms.Remarkably, our approach works in state-only settings without expert action labels, a setting which behavior cloning (BC) cannot solve.Empirical results demonstrate that our method learns from as few as a single expert demonstration and achieves improved performance on various control tasks.

1937. Fugatto 1: Foundational Generative Audio Transformer Opus 1

链接：https://iclr.cc/virtual/2025/poster/30598 abstract： Fugatto is a versatile audio synthesis and transformation model capable of following free-form text instructions with optional audio inputs. While large language models (LLMs) trained with text on a simple next-token prediction objective can learn to infer instructions directly from the data, models trained solely on audio data lack this capacity. This is because audio data does not inherently contain the instructions that were used to generate it. To overcome this challenge, we introduce a specialized dataset generation approach optimized for producing a wide range of audio generation and transformation tasks, ensuring the data reveals meaningful relationships between audio and language. Another challenge lies in achieving compositional abilities -- such as combining, interpolating between, or negating instructions -- using data alone. To address it, we propose ComposableART, an inference-time technique that extends classifier-free guidance to compositional guidance. It enables the seamless and flexible composition of instructions, leading to highly customizable audio outputs outside the training distribution. Our evaluations across a diverse set of tasks demonstrate that Fugatto performs competitively with specialized models, while ComposableART enhances its sonic palette and control over synthesis. Most notably, we highlight our framework's ability to execute emergent sounds and tasks -- sonic phenomena that transcend conventional audio generation -- unlocking new creative possibilities. \href{https://fugatto.github.io/}{Demo Website.}

1938. Gumbel Counterfactual Generation From Language Models

链接：https://iclr.cc/virtual/2025/poster/29541 abstract： Understanding and manipulating the causal generation mechanisms in language models is essential for controlling their behavior. Previous work has primarily relied on techniques such as representation surgery---e.g., model ablations or manipulation of linear subspaces tied to specific concepts---to intervene on these models. To understand the impact of interventions precisely, it is useful to examine counterfactuals---e.g., how a given sentence would have appeared had it been generated by the model following a specific intervention. We highlight that counterfactual reasoning is conceptually distinct from interventions, as articulated in Pearl's causal hierarchy. Based on this observation, we propose a framework for generating true string counterfactuals by reformulating language models as a structural equation model using the Gumbel-max trick, which we called Gumbel counterfactual generation. This reformulation allows us to model the joint distribution over original strings and their counterfactuals resulting from the same instantiation of the sampling noise. We develop an algorithm based on hindsight Gumbel sampling that allows us to infer the latent noise variables and generate counterfactuals of observed strings. Our experiments demonstrate that the approach produces meaningful counterfactuals while at the same time showing that commonly used intervention techniques have considerable undesired side effects.

1939. Linear Representations of Political Perspective Emerge in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28160 abstract： Large language models (LLMs) have demonstrated the ability to generate text that realistically reflects a range of different subjective human perspectives. This paper studies how LLMs are seemingly able to reflect more liberal versus more conservative viewpoints among other political perspectives in American politics. We show that LLMs possess linear representations of political perspectives within activation space, wherein more similar perspectives are represented closer together. To do so, we probe the attention heads across the layers of three open

transformer-based LLMs (Llama-2-7b-chat, Mistral-7b-instruct, Vicuna-7b). We first prompt models to generate text from the perspectives of different U.S. lawmakers. We then identify sets of attention heads whose activations linearly predict those lawmakers' DW-NOMINATE scores, a widely-used and validated measure of political ideology. We find that highly predictive heads are primarily located in the middle layers, often speculated to encode high-level concepts and tasks. Using probes only trained to predict lawmakers' ideology, we then show that the same probes can predict measures of news outlets' slant from the activations of models prompted to simulate text from those news outlets. These linear probes allow us to visualize, interpret, and monitor ideological stances implicitly adopted by an LLM as it generates open-ended responses. Finally, we demonstrate that by applying linear interventions to these attention heads, we can steer the model outputs toward a more liberal or conservative stance. Overall, our research suggests that LLMs possess a high-level linear representation of American political ideology and that by leveraging recent advances in mechanistic interpretability, we can identify, monitor, and steer the subjective perspective underlying generated text.

1940. Looking Backward: Streaming Video-to-Video Translation with Feature Banks

链接：https://iclr.cc/virtual/2025/poster/32105 abstract： This paper introduces StreamV2V, a diffusion model that achieves real-time streaming video-to-video (V2V) translation with user prompts. Unlike prior V2V methods using batches to process limited frames, we opt to process frames in a streaming fashion, to support unlimited frames. At the heart of StreamV2V lies a backward-looking principle that relates the present to the past. This is realized by maintaining a feature bank, which archives information from past frames. For incoming frames, StreamV2V extends self-attention to include banked keys and values, and directly fuses similar past features into the output. The feature bank is continually updated by merging stored and new features, making it compact yet informative. StreamV2V stands out for its adaptability and efficiency, seamlessly integrating with image diffusion models without fine-tuning. It can run 20 FPS on one A100 GPU, being 15$\times$, 46$\times$, 108$\times$, and 158$\times$ faster than FlowVid, CoDeF, Rerender, and TokenFlow, respectively. Quantitative metrics and user studies confirm StreamV2V's exceptional ability to maintain temporal consistency.

1941. Why Does the Effective Context Length of LLMs Fall Short?

链接：https://iclr.cc/virtual/2025/poster/28906 abstract： Advancements in distributed training and efficient attention mechanisms have significantly expanded the context window sizes of large language models (LLMs). However, recent work reveals that the effective context lengths of open-source LLMs often fall short, typically not exceeding half of their training lengths. In this work, we attribute this limitation to the left-skewed frequency distribution of relative positions formed in LLMs pretraining and post-training stages, which impedes their ability to effectively gather distant information.To address this challenge, we introduce Shifted Rotray Position Embedding (STRING). STRING shifts well-trained positions to overwrite the original ineffective positions during inference, enhancing performance within their existing training lengths. Experimental results show that without additional training, STRING dramatically improves the performance of the latest large-scale models, such as Llama3.1 70B and Qwen2 72B, by over 10 points on popular long-context benchmarks RULER and InfiniteBench, establishing new state-of-the-art results for open-source LLMs. Compared to commercial models, Llama 3.1 70B with STRING even achieves better performance than GPT-4-128K and clearly surpasses Claude 2 and Kimi-chat.

1942. Weighted Multi-Prompt Learning with Description-free Large Language Model Distillation

链接：https://iclr.cc/virtual/2025/poster/29892 abstract： Recent advances in pre-trained Vision Language Models (VLM) have shown promising potential for effectively adapting to downstream tasks through prompt learning, without the need for additional annotated paired datasets.To supplement the text information in VLM trained on correlations with vision data, new approaches leveraging Large Language Models (LLM) in prompts have been proposed, enhancing robustness to unseen and diverse data.Existing methods typically extract text-based responses (i.e., descriptions) from LLM to incorporate into prompts; however, this approach suffers from high variability and low reliability.In this work, we propose Description-free Multi-prompt Learning(DeMul), a novel method that eliminates the process of extracting descriptions and instead directly distills knowledge from LLM into prompts.By adopting a description-free approach, prompts can encapsulate richer semantics while still being represented as continuous vectors for optimization, thereby eliminating the need for discrete pre-defined templates.Additionally, in a multi-prompt setting, we empirically demonstrate the potential of prompt weighting in reflecting the importance of different prompts during training.Experimental results show that our approach achieves superior performance across 11 recognition datasets.

1943. Dobi-SVD: Differentiable SVD for LLM Compression and Some New Perspectives

链接：https://iclr.cc/virtual/2025/poster/28553 abstract： Large language models (LLMs) have sparked a new wave of AI applications; however, their substantial computational costs and memory demands pose significant challenges to democratizing access to LLMs for a broader audience. Singular Value Decomposition (SVD), a technique studied for decades, offers a hardware-independent and flexibly tunable solution for LLM compression. In this paper, we present new directions using SVD: we first theoretically analyze the optimality of truncating weights and truncating activations, then we further identify three key

issues on SVD-based LLM compression, including (1) How can we determine the optimal truncation position for each weight matrix in LLMs? (2) How can we efficiently update the weight matrices based on truncation position? (3) How can we address the inherent "injection" nature that results in the information loss of the SVD? We propose an effective approach, Dobi-SVD, to tackle the three issues. First, we propose a differentiable truncation-value learning mechanism, along with gradient-robust backpropagation, enabling the model to adaptively find the optimal truncation positions. Next, we utilize the Eckart-Young-Mirsky theorem to derive a theoretically optimal weight update formula through rigorous mathematical analysis. Lastly, by observing and leveraging the quantization-friendly nature of matrices after SVD decomposition, we reconstruct a mapping between truncation positions and memory requirements, establishing a bijection from truncation positions to memory. Experimental results show that with a 40\% parameter-compression rate, our method achieves a perplexity of 9.07 on the Wikitext2 dataset with the compressed LLama-7B model, a 78.7\% improvement over the state-of-the-art SVD for LLM compression method. We emphasize that Dobi-SVD is the first to achieve such a high-ratio LLM compression with minimal performance drop. We also extend our Dobi-SVD to VLM compression, achieving a 20\% increase in throughput with minimal performance degradation. We hope that the inference speedup—up to 12.4x on 12GB NVIDIA Titan Xp GPUs and 3x on 80GB A100 GPUs for LLMs, and 1.2x on 80GB A100 GPUs for VLMs—will bring significant benefits to the broader community such as robotics.

1944. Pitfalls of Evidence-Based AI Policy

链接：https://iclr.cc/virtual/2025/poster/31363 abstract： Nations across the world are working to govern AI. However, from a technical perspective, the best way to do this is not yet clear. Meanwhile, recent debates over AI regulation have led to calls for “evidence-based AI policy” which emphasize holding regulatory action to a high evidentiary standard. Evidence is of irreplaceable value to policymaking. However, holding regulatory action to too high an evidentiary standard can lead to systematic neglect of certain risks. In historical policy debates (e.g., over tobacco ca. 1965 and fossil fuels ca. 1990) “evidence-based policy” rhetoric is also a well-precedented strategy to downplay the urgency of action, delay regulation, and protect industry interests. Here, we argue that if the goal is evidence-based AI policy, the first regulatory objective must be to actively facilitate the process of identifying, studying, and deliberating about AI risks. We discuss a set of 16 regulatory goals to facilitate this and show that the EU, UK, USA, Brazil, Canada, and China all have substantial opportunities to adopt further evidence-seeking policies.

1945. Think-on-Graph 2.0: Deep and Faithful Large Language Model Reasoning with Knowledge-guided Retrieval Augmented Generation

链接：https://iclr.cc/virtual/2025/poster/28367 abstract： Retrieval-augmented generation (RAG) has improved large language models (LLMs) by using knowledge retrieval to overcome knowledge deficiencies. However, current RAG methods often fall short of ensuring the depth and completeness of retrieved information, which is necessary for complex reasoning tasks. In this work, we introduce Think-on-Graph 2.0 (ToG-2), a hybrid RAG framework that iteratively retrieves information from both unstructured and structured knowledge sources in a tight-coupling manner. Specifically, ToG-2 leverages knowledge graphs (KGs) to link documents via entities, facilitating deep and knowledge-guided context retrieval. Simultaneously, it utilizes documents as entity contexts to achieve precise and efficient graph retrieval. ToG-2 alternates between graph retrieval and context retrieval to search for in-depth clues relevant to the question, enabling LLMs to generate answers.We conduct a series of well-designed experiments to highlight the following advantages of ToG-2: 1) ToG-2 tightly couples the processes of context retrieval and graph retrieval, deepening context retrieval via the KG while enabling reliable graph retrieval based on contexts; 2) it achieves deep and faithful reasoning in LLMs through an iterative knowledge retrieval process of collaboration between contexts and the KG; and 3) ToG-2 is training-free and plug-and-play compatible with various LLMs. Extensive experiments demonstrate that ToG-2 achieves overall state-of-the-art (SOTA) performance on 6 out of 7 knowledge-intensive datasets with GPT-3.5, and can elevate the performance of smaller models (e.g., LLAMA-2-13B) to the level of GPT-3.5’s direct reasoning. The source code is available on https://anonymous.4open.science/r/ToG2.

1946. Towards Understanding Text Hallucination of Diffusion Models via Local Generation Bias

链接：https://iclr.cc/virtual/2025/poster/29614 abstract： Score-based diffusion models have achieved incredible performance in generating realistic images, audio, and video data. While these models produce high-quality samples with impressive details, they often introduce unrealistic artifacts, such as distorted fingers or hallucinated texts with no meaning. This paper focuses on textual hallucinations, where diffusion models correctly generate individual symbols but assemble them in a nonsensical manner. Through experimental probing, we consistently observe that such phenomenon is attributed it to the network's local generation bias. Denoising networks tend to produce outputs that rely heavily on highly correlated local regions, particularly when different dimensions of the data distribution are nearly pairwise independent. This behavior leads to a generation process that decomposes the global distribution into separate, independent distributions for each symbol, ultimately failing to capture the global structure, including underlying grammar. Intriguingly, this bias persists across various denoising network architectures including MLP and transformers which have the structure to model global dependency. These findings also provide insights into understanding other types of hallucinations, extending beyond text, as a result of implicit biases in the denoising models. Additionally, we theoretically analyze the training dynamics for a specific case involving a two-layer MLP learning parity points on a hypercube, offering an explanation of its underlying mechanism.

1947. Model-Agnostic Knowledge Guided Correction for Improved Neural

Surrogate Rollout

链接：https://iclr.cc/virtual/2025/poster/31063 abstract： Modeling the evolution of physical systems is critical to many applications in science and engineering. As the evolution of these systems is governed by partial differential equations (PDEs), there are a number of computational simulations which resolve these systems with high accuracy. However, as these simulations incur high computational costs, they are infeasible to be employed for large-scale analysis. A popular alternative to simulators are neural network surrogates which are trained in a data-driven manner and are much more computationally efficient. However, these surrogate models suffer from high rollout error when used autoregressively, especially when confronted with training data paucity. Existing work proposes to improve surrogate rollout error by either including physical loss terms directly in the optimization of the model or incorporating computational simulators as `differentiable layers' in the neural network. Both of these approaches have their challenges, with physical loss functions suffering from slow convergence for stiff PDEs and simulator layers requiring gradients which are not always available, especially in legacy simulators. We propose the Hybrid PDE Predictor with Reinforcement Learning (HyPER) model: a model-agnostic, RL based, cost-aware model which combines a neural surrogate, RL decision model, and a physics simulator (with or without gradients) to reduce surrogate rollout error significantly. In addition to reducing in-distribution rollout error by 47%-78%, HyPER learns an intelligent policy that is adaptable to changing physical conditions and resistant to noise corruption. Code available at https://github.com/scailab/HyPER.

1948. Learning Splitting Heuristics in Divide-and-Conquer SAT Solvers with Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/27974 abstract： We propose RDC-SAT, a novel approach to optimize splitting heuristics in Divide-and-Conquer SAT solvers using deep reinforcement learning. Our method dynamically extracts features from the current solving state whenever a split is required. These features, such as learned clauses, variable activity scores, and clause LBD (Literal Block Distance) values, are represented as a graph. A GNN integrated with an Actor-Critic model processes this graph to determine the optimal split variable. Unlike traditional linear state transitions characterized by Markov processes, divide-and-conquer challenges involve tree-like state transitions. To address this, we developed a reinforcement learning environment based on the Painless framework that efficiently handles these transitions. Additionally, we designed different discounted reward functions for satisfiable and unsatisfiable SAT problems, capable of handling tree-like state transitions. We trained our model using the Decentralized Proximal Policy Optimization (DPPO) algorithm on phase transition random 3-SAT problems and implemented the RDC-SAT solver, which operates in both GPU-accelerated and non-GPU modes. Evaluations show that RDC-SAT significantly improves the performance of D\&C solvers on phase transition random 3-SAT datasets and generalizes well to the SAT Competition 2023 dataset, substantially outperforming traditional splitting heuristics.

1949. Singular Subspace Perturbation Bounds via Rectangular Random Matrix Diffusions

链接：https://iclr.cc/virtual/2025/poster/30296 abstract： Given a matrix $A \in \mathbb{R}^{m\times d}$ with singular values $\sigma_1\geq \cdots \geq \sigma_d$, and a random matrix $G \in \mathbb{R}^{m\times d}$ with iid $N(0,T)$ entries for some $T>0$, we derive new bounds on the Frobenius distance between subspaces spanned by the top-k (right) singular vectors of A and $A+G$. This problem arises in numerous applications in statistics where a data matrix may be corrupted by Gaussian noise, and in the analysis of the Gaussian mechanism in differential privacy, where Gaussian noise is added to data to preserve private information. We show that, for matrices A where the gaps in the top-k singular values are roughly $\Omega(\sigma_k-\sigma_{k+1})$ the expected Frobenius distance between the subspaces is $\tilde{O}(\frac{\sqrt{d}}{\sigma_k-\sigma_{k+1}} \times \sqrt{T})$, improving on previous bounds by a factor of $\frac{\sqrt{m}}{\sqrt{d}}$. To obtain our bounds we view the perturbation to the singular vectors as a diffusion process-- the Dyson-Bessel process-- and use tools from stochastic calculus to track the evolution of the subspace spanned by the top-k singular vectors, which may be of independent interest.

1950. Decoding Game: On Minimax Optimality of Heuristic Text Generation Strategies

链接：https://iclr.cc/virtual/2025/poster/29351 abstract： Decoding strategies play a pivotal role in text generation for modern language models, yet a puzzling gap divides theory and practice. Surprisingly, strategies that should intuitively be optimal, such as Maximum a Posteriori (MAP), often perform poorly in practice. Meanwhile, popular heuristic approaches like Top-k and Nucleus sampling, which employ truncation and normalization of the conditional next-token probabilities, have achieved great empirical success but lack theoretical justifications. In this paper, we propose Decoding Game, a comprehensive theoretical framework which reimagines text generation as a two-player zero-sum game between Strategist, who seeks to produce text credible in the true distribution, and Nature, who distorts the true distribution adversarially. After discussing the decomposibility of multi-step generation, we derive the optimal strategy in closed form for one-step Decoding Game. It is shown that the adversarial Nature imposes an implicit regularization on likelihood maximization, and truncation-normalization methods are first-order approximations to the optimal strategy under this regularization. Additionally, by generalizing the objective and parameters of Decoding Game, near-optimal strategies encompass diverse methods such as greedy search, temperature scaling, and hybrids thereof. Numerical experiments are conducted to complement our theoretical analysis.

1951. Leveraging Variable Sparsity to Refine Pareto Stationarity in Multi-Objective Optimization

链接：https://iclr.cc/virtual/2025/poster/30551 abstract： Gradient-based multi-objective optimization (MOO) is essential in modern machine learning, with applications in e.g., multi-task learning, federated learning, algorithmic fairness and reinforcement learning. In this work, we first reveal some limitations of Pareto stationarity, a widely accepted first-order condition for Pareto optimality, in the presence of sparse function-variable structures. Next, to account for such sparsity, we propose a novel solution concept termed Refined Pareto Stationarity (RPS), which we prove is always sandwiched between Pareto optimality and Pareto stationarity. We give an efficient partitioning algorithm to automatically mine the function-variable dependency and substantially trim non-optimal Pareto stationary solutions. Then, we show that gradient-based descent algorithms in MOO can be enhanced with our refined partitioning. In particular, we propose Multiple Gradient Descent Algorithm with Refined Partition (RP-MGDA) as an example method that converges to RPS, while still enjoying a similar per-step complexity and convergence rate. Lastly, we validate our approach through experiments on both synthetic examples and realistic application scenarios where distinct function-variable dependency structures appear. Our results highlight the importance of exploiting function-variable structure in gradient-based MOO, and provide a seamless enhancement to existing approaches.

1952. Stochastic Bandits Robust to Adversarial Attacks

链接：https://iclr.cc/virtual/2025/poster/27916 abstract： This paper investigates stochastic multi-armed bandit algorithms that are robust to adversarial attacks, where an attacker can first observe the learner's action and *then* alter their reward observation.We study two cases of this model, with or without the knowledge of an attack budget C, defined as an upper bound of the summation of the difference between the actual and altered rewards. For both cases, we devise two types of algorithms with regret bounds having additive or multiplicative C dependence terms.For the known attack budget case, we prove our algorithms achieve the regret bound of ${O}((K/\Delta)\log T + KC)$ and $\tilde{O}(\sqrt{KTC})$ for the additive and multiplicative C terms, respectively, where K is the number of arms, T is the time horizon, Δ is the gap between the expected rewards of the optimal arm and the second-best arm, and $\tilde{O}$ hides the logarithmic factors.For the unknown case, we prove our algorithms achieve the regret bound of $\tilde{O}(\sqrt{KT} + KC^2)$ and $\tilde{O}(KC\sqrt{T})$ for the additive and multiplicative C terms, respectively.In addition to these upper bound results, we provide several lower bounds showing the tightness of our bounds and the optimality of our algorithms.These results delineate an intrinsic separation between the bandits with attacks and corruption models.

1953. Model-Free Offline Reinforcement Learning with Enhanced Robustness

链接：https://iclr.cc/virtual/2025/poster/29669 abstract： Offline reinforcement learning (RL) has gained considerable attention for its ability to learn policies from pre-collected data without real-time interaction, which makes it particularly useful for high-risk applications. However, due to its reliance on offline datasets, existing works inevitably introduce assumptions to ensure effective learning, which, however, often lead to a trade-off between robustness to model mismatch and scalability to large environments. In this paper, we enhance both aspects with a novel double-pessimism principle, which conservatively estimates performance and accounts for both limited data and potential model mismatches, two major reasons for the previous trade-off. We then propose a universal, model-free algorithm to learn a policy that is robust to potential environment mismatches, which enhances robustness in a scalable manner. Furthermore, we provide a sample complexity analysis of our algorithm when the mismatch is modeled by the l_α-norm, which also theoretically demonstrates the efficiency of our method. Extensive experiments further demonstrate that our approach significantly improves robustness in a more scalable manner than existing methods.

1954. Efficient stagewise pretraining via progressive subnetworks

链接：https://iclr.cc/virtual/2025/poster/29274 abstract： Recent developments in large language models have sparked interest in efficientpretraining methods. Stagewise training approaches to improve efficiency, likegradual stacking and layer dropping (Reddi et al., 2023; Zhang & He, 2020), haverecently garnered attention. The prevailing view suggests that stagewise droppingstrategies, such as layer dropping, are ineffective, especially when compared tostacking-based approaches. This paper challenges this notion by demonstratingthat, with proper design, dropping strategies can be competitive, if not better, thanstacking methods. Specifically, we develop a principled stagewise training framework, progressive subnetwork training, which only trains subnetworks within themodel and progressively increases the size of subnetworks during training, until ittrains the full network. We propose an instantiation of this framework — RandomPart Training (RAPTR) — that selects and trains only a random subnetwork (e.g.depth-wise, width-wise) of the network at each step, progressively increasing thesize in stages. We show that this approach not only generalizes prior works likelayer dropping but also fixes their key issues. Furthermore, we establish a theoretical basis for such approaches and provide justification for (a) increasing complexity of subnetworks in stages, conceptually diverging from prior works on layerdropping, and (b) stability in loss across stage transitions in presence of key modern architecture components like residual connections and layer norms. Throughcomprehensive experiments, we demonstrate that RAPTR can significantly speedup training of standard benchmarks like BERT and UL2, up to 33% compared tostandard training and, surprisingly, also shows better downstream performance onUL2, improving QA tasks and SuperGLUE by 1.5%; thereby, providing evidenceof better inductive bias.

1955. A Multi-Power Law for Loss Curve Prediction Across Learning Rate Schedules

链接：https://iclr.cc/virtual/2025/poster/30030 abstract： Training large models is both resource-intensive and time-consuming, making it crucial to understand the quantitative relationship between model performance and hyperparameters. In this paper, we derive an empirical law that predicts pretraining loss for large language models for every intermediate training step across various learning rate schedules, including constant, cosine, and step decay schedules. Our proposed law takes a multi-power form, combining a power law based on the sum of learning rates and additional power laws to account for a loss reduction effect as learning rate decays. We validate this law extensively on Llama-2 models of varying sizes and demonstrate that, after fitting on a few learning rate schedules, it accurately predicts the loss curves for unseen schedules of different shapes and horizons. Moreover, by minimizing the predicted final pretraining loss across learning rate schedules, we are able to find a schedule that outperforms the widely-used cosine learning rate schedule. Interestingly, this automatically discovered schedule bears some resemblance to the recently proposed Warmup-Stable-Decay (WSD) schedule (Hu et al, 2024) but achieves a slightly lower final loss. We believe these results could offer valuable insights for understanding the dynamics of pretraining and for designing learning rate schedules to improve efficiency.

1956. SONICS: Synthetic Or Not - Identifying Counterfeit Songs

链接：https://iclr.cc/virtual/2025/poster/32089 abstract： The recent surge in AI-generated songs presents exciting possibilities and challenges. These innovations necessitate the ability to distinguish between human-composed and synthetic songs to safeguard artistic integrity and protect human musical artistry. Existing research and datasets in fake song detection only focus on singing voice deepfake detection (SVDD), where the vocals are AI-generated but the instrumental music is sourced from real songs. However, these approaches are inadequate for detecting contemporary end-to-end artificial songs where all components (vocals, music, lyrics, and style) could be AI-generated. Additionally, existing datasets lack music-lyrics diversity, long-duration songs, and open-access fake songs. To address these gaps, we introduce SONICS, a novel dataset for end-to-end Synthetic Song Detection (SSD), comprising over 97k songs (4,751 hours) with over 49k synthetic songs from popular platforms like Suno and Udio. Furthermore, we highlight the importance of modeling long-range temporal dependencies in songs for effective authenticity detection, an aspect entirely overlooked in existing methods. To utilize long-range patterns, we introduce SpecTTTra, a novel architecture that significantly improves time and memory efficiency over conventional CNN and Transformer-based models. For long songs, our top-performing variant outperforms ViT by 8% in F1 score, is 38% faster, and uses 26% less memory, while also surpassing ConvNeXt with a 1% F1 score gain, 20% speed boost, and 67% memory reduction.

1957. Behavioral Entropy-Guided Dataset Generation for Offline Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29971 abstract： Entropy-based objectives are widely used to perform state space exploration in reinforcement learning (RL) and dataset generation for offline RL. Behavioral entropy (BE), a rigorous generalization of classical entropies that incorporates cognitive and perceptual biases of agents, was recently proposed for discrete settings and shown to be a promising metric for robotic exploration problems. In this work, we propose using BE as a principled exploration objective for systematically generating datasets that provide diverse state space coverage in complex, continuous, potentially high-dimensional domains. To achieve this, we extend the notion of BE to continuous settings, derive tractable k-nearest neighbor estimators, provide theoretical guarantees for these estimators, and develop practical reward functions that can be used with standard RL methods to learn BE-maximizing policies. Using standard MuJoCo environments, we experimentally compare the performance of offline RL algorithms for a variety of downstream tasks on datasets generated using BE, R\'{e}nyi, and Shannon entropy-maximizing policies, as well as the SMM and RND algorithms. We find that offline RL algorithms trained on datasets collected using BE outperform those trained on datasets collected using Shannon entropy, SMM, and RND on all tasks considered, and on 80\% of the tasks compared to datasets collected using Renyi entropy.

1958. Learned Reference-based Diffusion Sampler for multi-modal distributions

链接：https://iclr.cc/virtual/2025/poster/28860 abstract： Over the past few years, several approaches utilizing score-based diffusion have been proposed to sample from probability distributions, that is without having access to exact samples and relying solely on evaluations of unnormalized densities. The resulting samplers approximate the time-reversal of a noising diffusion process, bridging the target distribution to an easy-to-sample base distribution. In practice, the performance of these methods heavily depends on key hyperparameters that require ground truth samples to be accurately tuned. Our work aims to highlight and address this fundamental issue, focusing in particular on multi-modal distributions, which pose significant challenges for existing sampling methods. Building on existing approaches, we introduce Learned Reference-based Diffusion Sampler (LRDS), a methodology specifically designed to leverage prior knowledge on the location of the target modes in order to bypass the obstacle of hyperparameter tuning. LRDS proceeds in two steps by (i) learning a reference diffusion model on samples located in high-density space regions and tailored for multimodality, and (ii) using this reference model to foster the training of a diffusion-based sampler. We experimentally demonstrate that LRDS best exploits prior knowledge on the target distribution compared to competing algorithms on a variety of challenging distributions.

1959. PolaFormer: Polarity-aware Linear Attention for Vision Transformers

链接：https://iclr.cc/virtual/2025/poster/28592 abstract： Linear attention has emerged as a promising alternative to softmax-based attention, leveraging kernelized feature maps to reduce complexity from quadratic to linear in sequence length. However, the non-negative constraint on feature maps and the relaxed exponential function used in approximation lead to significant information loss compared to the original query-key dot products, resulting in less discriminative attention maps with higher entropy. To address the missing interactions driven by negative values in query-key pairs, we propose a polarity-aware linear attention mechanism that explicitly models both same-signed and opposite-signed query-key interactions, ensuring comprehensive coverage of relational information. Furthermore, to restore the spiky properties of attention maps, we provide a theoretical analysis proving the existence of a class of element-wise functions (with positive first and second derivatives) that can reduce entropy in the attention distribution. For simplicity, and recognizing the distinct contributions of each dimension, we employ a learnable power function for rescaling, allowing strong and weak attention signals to be effectively separated. Extensive experiments demonstrate that the proposed PolaFormer improves performance on various vision tasks, enhancing both expressiveness and efficiency by up to 4.6%.

1960. BioDiscoveryAgent: An AI Agent for Designing Genetic Perturbation Experiments

链接：https://iclr.cc/virtual/2025/poster/30240 abstract： Agents based on large language models have shown great potential in accelerating scientific discovery by leveraging their rich background knowledge and reasoning capabilities. In this paper, we introduce BioDiscoveryAgent, an agent that designs new experiments, reasons about their outcomes, and efficiently navigates the hypothesis space to reach desired solutions. We demonstrate our agent on the problem of designing genetic perturbation experiments, where the aim is to find a small subset out of many possible genes that, when perturbed, result in a specific phenotype (e.g., cell growth). Utilizing its biological knowledge, BioDiscoveryAgent can uniquely design new experiments without the need to train a machine learning model or explicitly design an acquisition function as in Bayesian optimization. Moreover, BioDiscoveryAgent using Claude 3.5 Sonnet achieves an average of 21% improvement in predicting relevant genetic perturbations across six datasets, and a 46% improvement in the harder task of non-essential gene perturbation, compared to existing Bayesian optimization baselines specifically trained for this task. Our evaluation includes one dataset that is unpublished, ensuring it is not part of the language model's training data. Additionally, BioDiscoveryAgent predicts gene combinations to perturb more than twice as accurately as a random baseline, a task so far not explored in the context of closed-loop experiment design. The agent also has access to tools for searching the biomedical literature, executing code to analyze biological datasets, and prompting another agent to critically evaluate its predictions. Overall, BioDiscoveryAgent is interpretable at every stage, representing an accessible new paradigm in the computational design of biological experiments with the potential to augment scientists' efficacy.

1961. Effective Interplay between Sparsity and Quantization: From Theory to Practice

链接：https://iclr.cc/virtual/2025/poster/27855 abstract： The increasing size of deep neural networks (DNNs) necessitates effective model compression to reduce their computational and memory footprints. Sparsity and quantization are two prominent compression methods that have been shown to reduce DNNs' computational and memory footprints significantly while preserving model accuracy. However, how these two methods interact when combined together remains a key question for developers, as many tacitly assume that they are orthogonal, meaning that their combined use does not introduce additional errors beyond those introduced by each method independently. In this paper, we provide the first mathematical proof that sparsity and quantization are non-orthogonal. We corroborate these results with experiments spanning a range of large language models, including the OPT and LLaMA model families (with 125M to 8B parameters), and vision models like ViT and ResNet. We show that the order in which we apply these methods matters because applying quantization before sparsity may disrupt the relative importance of tensor elements, which may inadvertently remove significant elements from a tensor. More importantly, we show that even if applied in the correct order, the compounded errors from sparsity and quantization can significantly harm accuracy. Our findings extend to the efficient deployment of large models in resource-constrained compute platforms to reduce serving cost, offering insights into best practices for applying these compression methods to maximize hardware resource efficiency without compromising accuracy.

1962. On the Convergence of No-Regret Dynamics in Information Retrieval Games with Proportional Ranking Functions

链接：https://iclr.cc/virtual/2025/poster/28641 abstract： Publishers who publish their content on the web act strategically, in a behavior that can be modeled within the online learning framework. Regret, a central concept in machine learning, serves as a canonical measure for assessing the performance of learning agents within this framework.We prove that any proportional content ranking function with a concave activation function induces games in which no-regret learning dynamics converge. Moreover, for proportional ranking functions, we prove the equivalence of the concavity of the activation function, the social concavity of the induced games and the concavity of the induced games.We also study the empirical trade-offs between publishers' and users' welfare, under different choices of the activation function, using a state-of-the-art no-regret dynamics algorithm. Furthermore, we demonstrate how the choice of the ranking function and changes in the ecosystem structure affect

these welfare measures, as well as the dynamics' convergence rate.

1963. Diffusion Models as Cartoonists: The Curious Case of High Density Regions

链接：https://iclr.cc/virtual/2025/poster/29641 abstract： We investigate what kind of images lie in the high-density regions of diffusion models. We introduce a theoretical mode-tracking process capable of pinpointing the exact mode of the denoising distribution, and we propose a practical high-density sampler that consistently generates images of higher likelihood than usual samplers. Our empirical findings reveal the existence of significantly higher likelihood samples that typical samplers do not produce, often manifesting as cartoon-like drawings or blurry images depending on the noise level. Curiously, these patterns emerge in datasets devoid of such examples. We also present a novel approach to track sample likelihoods in diffusion SDEs, which remarkably incurs no additional computational cost. Code is available at https://github.com/Aalto-QuML/high-density-diffusion

1964. Fine-Tuning Attention Modules Only: Enhancing Weight Disentanglement in Task Arithmetic

链接：https://iclr.cc/virtual/2025/poster/28970 abstract： In recent years, task arithmetic has garnered increasing attention. This approach edits pre-trained models directly in weight space by combining the fine-tuned weights of various tasks into a unified model. Its efficiency and cost-effectiveness stem from its training-free combination, contrasting with traditional methods that require model training on large datasets for multiple tasks. However, applying such a unified model to individual tasks can lead to interference from other tasks (lack of weight disentanglement). To address this issue, Neural Tangent Kernel (NTK) linearization has been employed to leverage a "kernel behavior", facilitating weight disentanglement and mitigating adverse effects from unrelated tasks. Despite its benefits, NTK linearization presents drawbacks, including doubled training costs, as well as reduced performance of individual models. To tackle this problem, we propose a simple yet effective and efficient method that is to finetune the attention modules only in the Transformer. Our study reveals that the attention modules exhibit kernel behavior, and fine-tuning the attention modules only significantly improves weight disentanglement. To further understand how our method improves the weight disentanglement of task arithmetic, we present a comprehensive study of task arithmetic by differentiating the role of the representation module and task-specific module. In particular, we find that the representation module plays an important role in improving weight disentanglement whereas the task-specific modules such as the classification heads can degenerate the weight disentanglement performance.

1965. ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities

链接：https://iclr.cc/virtual/2025/poster/28507 abstract： Forecasts of future events are essential inputs into informed decision-making. Machine learning (ML) systems have the potential to deliver forecasts at scale, but there is no framework for evaluating the accuracy of ML systems on a standardized set of forecasting questions. To address this gap, we introduce ForecastBench: a dynamic benchmark that evaluates the accuracy of ML systems on an automatically generated and regularly updated set of 1,000 forecasting questions. To avoid any possibility of data leakage, ForecastBench is comprised solely of questions about future events that have no known answer at the time of submission. We quantify the capabilities of current ML systems by collecting forecasts from expert (human) forecasters, the general public, and LLMs on a random subset of questions from the benchmark ($N=200$). While LLMs have achieved super-human performance on many benchmarks, they perform less well here: expert forecasters outperform the top-performing LLM (p-value <0.001). We display system and human scores in a public leaderboard at www.forecastbench.org.

1966. Generalization and Distributed Learning of GFlowNets

链接：https://iclr.cc/virtual/2025/poster/29760 abstract： Conventional wisdom attributes the success of Generative Flow Networks (GFlowNets) to their ability to exploit the compositional structure of the sample space for learning generalizable flow functions (Bengio et al., 2021). Despite the abundance of empirical evidence, formalizing this belief with verifiable non-vacuous statistical guarantees has remained elusive. We address this issue with the first data-dependent generalization bounds for GFlowNets. We also elucidate the negative impact of the state space size on the generalization performance of these models via Azuma-Hoeffding-type oracle PAC-Bayesian inequalities. We leverage our theoretical insights to design a novel distributed learning algorithm for GFlowNets, which we call Subgraph Asynchronous Learning (SAL). In a nutshell, SAL utilizes a divide-and-conquer strategy: multiple GFlowNets are trained in parallel on smaller subnetworks of the flow network, and then aggregated with an additional GFlowNet that allocates appropriate flow to each subnetwork. Our experiments with synthetic and real-world problems demonstrate the benefits of SAL over centralized training in terms of mode coverage and distribution matching.

1967. ToolGen: Unified Tool Retrieval and Calling via Generation

链接：https://iclr.cc/virtual/2025/poster/29311 abstract： As large language models (LLMs) advance, their inability to autonomously execute tasks by directly interacting with external tools remains a critical limitation. Traditional methods rely on inputting tool descriptions as context, which is constrained by context length and requires separate, often inefficient, retrieval mechanisms. We introduce ToolGen, a paradigm shift that integrates tool knowledge directly into the LLM’s parameters by

representing each tool as a unique token. This enables the LLM to generate tool calls and arguments as part of its next token prediction capabilities, seamlessly blending tool invocation with language generation. Our framework allows the LLM to access and utilize a vast amount of tools with no additional retrieval step, significantly enhancing both performance and scalability. Experimental results with over 47,000 tools show that ToolGen not only achieves superior results in both tool retrieval and autonomous task completion but also sets the stage for a new era of AI agents that can adapt to tools across diverse domains. By fundamentally transforming tool retrieval into a generative process, ToolGen paves the way for more versatile, efficient, and autonomous AI systems. ToolGen enables end-to-end tool learning and opens opportunities for integration with other advanced techniques such as chain-of-thought and reinforcement learning, thereby expanding the practical capabilities of LLMs

1968. Differentiation and Specialization of Attention Heads via the Refined Local Learning Coefficient

链接：https://iclr.cc/virtual/2025/poster/29600 abstract： We introduce refined variants of the Local Learning Coefficient (LLC), a measure of model complexity grounded in singular learning theory, to study the development of internal structure in transformer language models during training. By applying these refined LLCs (rLLCs) to individual components of a two-layer attention-only transformer, we gain novel insights into the progressive differentiation and specialization of attention heads. Our methodology reveals how attention heads differentiate into distinct functional roles over the course of training, analyzes the types of data these heads specialize to process, and discovers a previously unidentified multigram circuit. These findings demonstrate that rLLCs provide a principled, quantitative toolkit for developmental interpretability, which aims to understand models through their evolution across the learning process. This work advances the field of developmental interpretability by providing a mathematically rigorous approach to understanding neural networks through the lens of their learning process. More broadly, this work takes a step towards establishing the correspondence between data distributional structure, geometric properties of the loss landscape, learning dynamics, and emergent computational structures in neural networks.

1969. WardropNet: Traffic Flow Predictions via Equilibrium-Augmented Learning

链接：https://iclr.cc/virtual/2025/poster/30834 abstract： When optimizing transportation systems, anticipating traffic flows is a central element. Yet, computing such traffic equilibria remains computationally expensive. Against this background, we introduce a novel combinatorial optimization augmented neural network pipeline that allows for fast and accurate traffic flow predictions. We propose WardropNet, a neural network that combines classical layers with a subsequent equilibrium layer: the first ones inform the latter by predicting the parameterization of the equilibrium problem's latency functions. Using supervised learning we minimize the difference between the actual traffic flow and the predicted output. We show how to leverage a Bregman divergence fitting the geometry of the equilibria, which allows for end-to-end learning. WardropNet outperforms pure learning-based approaches in predicting traffic equilibria for realistic and stylized traffic scenarios. On realistic scenarios, WardropNet improves on average for time-invariant predictions by up to 72\% and for time-variant predictions by up to 23\% over pure learning-based approaches.

1970. When do GFlowNets learn the right distribution?

链接：https://iclr.cc/virtual/2025/poster/30708 abstract： Generative Flow Networks (GFlowNets) are an emerging class of sampling methods for distributions over discrete and compositional objects, e.g., graphs. In spite of their remarkable success in problems such as drug discovery and phylogenetic inference, the question of when and whether GFlowNets learn to sample from the target distribution remains underexplored. To tackle this issue, we first assess the extent to which a violation of the detailed balance of the underlying flow network might hamper the correctness of GFlowNet's sampling distribution. In particular, we demonstrate that the impact of an imbalanced edge on the model's accuracy is influenced by the total amount of flow passing through it and, as a consequence, is unevenly distributed across the network. We also argue that, depending on the parameterization, imbalance may be inevitable. In this regard, we consider the problem of sampling from distributions over graphs with GFlowNets parameterized by graph neural networks (GNNs) and show that the representation limits of GNNs delineate which distributions these GFlowNets can approximate. Lastly, we address these limitations by proposing a theoretically sound and computationally tractable metric for assessing GFlowNets, experimentally showing it is a better proxy for correctness than popular evaluation protocols.

1971. CoRNStack: High-Quality Contrastive Data for Better Code Retrieval and Reranking

链接：https://iclr.cc/virtual/2025/poster/28662 abstract： Effective code retrieval plays a crucial role in advancing code generation, bug fixing, and software maintenance, particularly as software systems increase in complexity. While current code embedding models have demonstrated promise in retrieving code snippets for small-scale, well-defined tasks, they often underperform in more demanding real-world applications such as bug localization within GitHub repositories. We hypothesize that a key issue is their reliance on noisy and inconsistent datasets for training, which impedes their ability to generalize to more complex retrieval scenarios. To address these limitations, we introduce CoRNStack, a large-scale, high-quality contrastive training dataset for code that spans multiple programming languages. This dataset is curated using consistency filtering to eliminate noisy positives and is further enriched with mined hard negatives, thereby facilitating more effective learning. We

demonstrate that contrastive training of embedding models using CoRNStack leads to state-of-the-art performance across a variety of code retrieval tasks. Furthermore, the dataset can be leveraged for training code reranking models, a largely underexplored area compared to text reranking. Our finetuned code reranking model significantly improves the ranking quality over the retrieved results. Finally, by employing our code retriever and reranker together, we demonstrate significant improvements in function localization for GitHub issues, an importantcomponent of real-world software development.

1972. Equivariant Denoisers Cannot Copy Graphs: Align Your Graph Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/28333 abstract： Graph diffusion models, dominant in graph generative modeling, remain underexplored for graph-to-graph translation tasks like chemical reaction prediction. We demonstrate that standard permutation equivariant denoisers face fundamental limitations in these tasks due to their inability to break symmetries in noisy inputs. To address this, we propose \emph{aligning} input and target graphs to break input symmetries while preserving permutation equivariance in non-matching graph portions. Using retrosynthesis (i.e., the task of predicting precursors for synthesis of a given target molecule) as our application domain, we show how alignment dramatically improves discrete diffusion model performance from 5\% to a SOTA-matching 54.7\% top-1 accuracy. Code is available at https://github.com/Aalto-QuML/DiffAlign.

1973. Towards a Complete Logical Framework for GNN Expressiveness

链接：https://iclr.cc/virtual/2025/poster/28279 abstract： Designing expressive Graph neural networks (GNNs) is an important topic in graph machine learning fields. Traditionally, the Weisfeiler-Lehman (WL) test has been the primary measure for evaluating GNN expressiveness. However, high-order WL tests can be obscure, making it challenging to discern the specific graph patterns captured by them. Given the connection between WL tests and first-order logic, some have explored the logical expressiveness of Message Passing Neural Networks. This paper aims to establish a comprehensive and systematic relationship between GNNs and logic. We propose a framework for identifying the equivalent logical formulas for arbitrary GNN architectures, which not only explains existing models, but also provides inspiration for future research. As case studies, we analyze multiple classes of prominent GNNs within this framework, unifying different subareas of the field. Additionally, we conduct a detailed examination of homomorphism expressivity from a logical perspective and present a general method for determining the homomorphism expressivity of arbitrary GNN models, as well as addressing several open problems.

1974. Improving Semantic Understanding in Speech Language Models via Brain-tuning

链接：https://iclr.cc/virtual/2025/poster/30063 abstract： Speech language models align with human brain responses to natural language to an impressive degree. However, current models rely heavily on low-level speech features, indicating they lack brain-relevant semantics which limits their utility as model organisms of semantic processing in the brain. In this work, we address this limitation by inducing brain-relevant bias directly into the models via fine-tuning with fMRI recordings of people listening to natural stories--a process we name brain-tuning. After testing it on 3 different pretrained model families, we show that brain-tuning not only improves overall alignment with new brain recordings in semantic language regions, but also reduces the reliance on low-level speech features for this alignment. Excitingly, we further show that brain-tuning leads to 1) consistent improvements in performance on semantic downstream tasks and 2) a representational space with increased semantic preference. Our results provide converging evidence, for the first time, that incorporating brain signals into the training of language models improves the models' semantic understanding.

1975. Perplexed by Perplexity: Perplexity-Based Data Pruning With Small Reference Models

链接：https://iclr.cc/virtual/2025/poster/31214 abstract： In this work, we investigate whether small language models can determine high-quality subsets of large-scale text datasets that improve the performance of larger language models. While existing work has shown that pruning based on the perplexity of a larger model can yield high-quality data, we investigate whether smaller models can be used for perplexity-based pruning and how pruning is affected by the domain composition of the data being pruned. We demonstrate that for multiple dataset compositions, perplexity-based pruning of pretraining data can significantly improve downstream task performance: pruning based on perplexities computed with a 125 million parameter model improves the average performance on downstream tasks of a 3 billion parameter model by up to 2.04 and achieves up to a 1.45× reduction in pretraining steps to reach commensurate baseline performance. Furthermore, we demonstrate that such perplexity-based data pruning also yields downstream performance gains in the over-trained and data-constrained regimes.

1976. ROUTE: Robust Multitask Tuning and Collaboration for Text-to-SQL

链接：https://iclr.cc/virtual/2025/poster/30587 abstract： Despite the significant advancements in Text-to-SQL (Text2SQL) facilitated by large language models (LLMs), the latest state-of-the-art techniques are still trapped in the in-context learning of closed-source LLMs (e.g., GPT-4), which limits their applicability in open scenarios. To address this challenge, we propose a novel RObust mUltitask Tuning and collaboration mEthod (ROUTE) to improve the comprehensive capabilities of open-source

LLMs for Text2SQL, thereby providing a more practical solution. Our approach begins with multi-task supervised fine-tuning (SFT) using various synthetic training data related to SQL generation. Unlike existing SFT-based Text2SQL methods, we introduced several additional SFT tasks, including schema linking, noise correction, and continuation writing. Engaging in a variety of SQL generation tasks enhances the model's understanding of SQL syntax and improves its ability to generate high-quality SQL queries. Additionally, inspired by the collaborative modes of LLM agents, we introduce a Multitask Collaboration Prompting (MCP) strategy. This strategy leverages collaboration across several SQL-related tasks to reduce hallucinations during SQL generation, thereby maximizing the potential of enhancing Text2SQL performance through explicit multitask capabilities. Extensive experiments and in-depth analyses have been performed on eight open-source LLMs and five widely-used benchmarks. The results demonstrate that our proposal outperforms the latest Text2SQL methods and yields leading performance.

1977. Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

链接：https://iclr.cc/virtual/2025/poster/30482 abstract： Vision-language models (VLMs) have revolutionized machine learning by leveraging large pre-trained models to tackle various downstream tasks. Although label, training, and data efficiency have improved, many state-of-the-art VLMs still require task-specific hyperparameter tuning and fail to fully exploit test samples. To overcome these challenges, we propose a graph-based approach for label-efficient adaptation and inference. Our method dynamically constructs a graph over text prompts, few-shot examples, and test samples, using label propagation for inference without task-specific tuning. Unlike existing zero-shot label propagation techniques, our approach requires no additional unlabeled support set and effectively leverages the test sample manifold through dynamic graph expansion. We further introduce a context-aware feature re-weighting mechanism to improve task adaptation accuracy. Additionally, our method supports efficient graph expansion, enabling real-time inductive inference. Extensive evaluations on downstream tasks, such as fine-grained categorization and out-of-distribution generalization, demonstrate the effectiveness of our approach. The source code is available at https://github.com/Yushu-Li/ECALP.

1978. Scaling Laws for Precision

链接：https://iclr.cc/virtual/2025/poster/27832 abstract： Low precision training and inference affect both the quality and cost of language models, but current scaling laws do not account for this. In this work, we devise "precision-aware" scaling laws for both training and inference. We propose that training in lower precision reduces the model's "effective parameter count," allowing us to predict the additional loss incurred from training in low precision and post-train quantization. For inference, we find that the degradation introduced by post-training quantization increases as models are trained on more data, eventually making additional pretraining data actively harmful. For training, our scaling laws allow us to predict the loss of a model with different parts in different precisions, and suggest that training larger models in lower precision can be compute optimal. We unify the scaling laws for post and pretraining quantization to arrive at a single functional form that predicts degradation from training and inference in varied precisions. We fit on over 465 pretraining runs and validate our predictions on model sizes up to 1.7B parameters trained on up to 26B tokens.

1979. Human-inspired Episodic Memory for Infinite Context LLMs

链接：https://iclr.cc/virtual/2025/poster/30579 abstract： Large language models (LLMs) have shown remarkable capabilities, but still struggle with processing extensive contexts, limiting their ability to maintain coherence and accuracy over long sequences. In contrast, the human brain excels at organising and retrieving episodic experiences across vast temporal scales, spanning a lifetime. In this work, we introduce EM-LLM, a novel approach that integrates key aspects of human episodic memory and event cognition into LLMs with no fine-tuning, enabling them to handle practically infinite context lengths while maintaining computational efficiency. EM-LLM organises sequences of tokens into coherent episodic events using a combination of Bayesian surprise and graph-theoretic boundary refinement in an online fashion. When needed, these events are retrieved through a two-stage memory process, combining similarity-based and temporally contiguous retrieval for efficient, human-inspired access to relevant information. Experiments on the LongBench and ∞-Bench benchmarks demonstrate EM-LLM's superior performance, consistently outperforming the state-of-the-art retrieval model InfLLM across various baseline LLMs. In addition, EM-LLM outperforms its popular counterpart, RAG, in a wide range of tasks, while requiring similar resources. Notably, EM-LLM's performance even surpasses full-context models in most tasks, while successfully performing retrieval across 10 million tokens -- a scale computationally infeasible for such models. Finally, our analysis reveals strong correlations between EM-LLM's event segmentation and human-perceived events, suggesting parallels between this artificial system and its biological counterpart, thereby offering a novel computational framework for exploring human memory mechanisms.

1980. Rethinking Shapley Value for Negative Interactions in Non-convex Games

链接：https://iclr.cc/virtual/2025/poster/29127 abstract： We study causal interactions for payoff allocation in cooperative game theory, including quantifying feature attribution for deep learning models. Most feature attribution methods mainly stem from the criteria of the Shapley value, which assigns fair payoffs to players based on their expected contribution in a cooperative game. However, interactions between players in the game do not explicitly appear in the original formulation of the Shapley value. In this work, we reformulate the Shapley value to clarify the role of interactions and discuss implicit assumptions from a game-

theoretical perspective. Our theoretical analysis demonstrates that when negative interactions exist—common in deep learning models—the efficiency axiom can lead to the undervaluation of attributions or payoffs. We suggest a new allocation rule that decomposes contributions into interactions and aggregates positive parts for non-convex games. Furthermore, we propose an approximation algorithm to reduce the cost of interaction computation which can be applied to differentiable functions such as deep learning models. Our approach mitigates counterintuitive attribution outcomes observed in existing methods, ensuring that features critical to a model's decision receive appropriate attribution.

1981. MeteoRA: Multiple-tasks Embedded LoRA for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/27732 abstract： The pretrain+fine-tune paradigm is foundational for deploying large language models (LLMs) across various downstream applications. Within this framework, Low-Rank Adaptation (LoRA) stands out for its parameter-efficient fine-tuning (PEFT), producing numerous reusable task-specific LoRA adapters. However, this approach requires explicit task intention selection, posing challenges for autonomous task sensing and switching during inference with multiple existing LoRA adapters embedded in a single LLM. In this work, we introduce MeteoRA (Multiple-Tasks embedded LoRA), a scalable and efficient framework that reuses multiple task-specific LoRA adapters into the base LLM via a full-mode Mixture-of-Experts (MoE) architecture. This framework also includes novel MoE forward acceleration strategies to address the efficiency challenges of traditional MoE implementations. Our evaluation, using the LlaMA2-13B and LlaMA3-8B base models equipped with 28 existing LoRA adapters through MeteoRA, demonstrates equivalent performance with the traditional PEFT method. Moreover, the LLM equipped with MeteoRA achieves superior performance in handling composite tasks, effectively solving ten sequential problems in a single inference pass, thereby demonstrating the framework's enhanced capability for timely adapter switching.

1982. Matérn Kernels for Tunable Implicit Surface Reconstruction

链接：https://iclr.cc/virtual/2025/poster/29789 abstract： We propose to use the family of Matérn kernels for implicit surface reconstruction, building upon the recent success of kernel methods for 3D reconstruction of oriented point clouds. As we show from a theoretical and practical perspective, Matérn kernels have some appealing properties which make them particularly well suited for surface reconstruction---outperforming state-of-the-art methods based on the arc-cosine kernel while being significantly easier to implement, faster to compute, and scalable. Being stationary, we demonstrate that Matérn kernels allow for tunable surface reconstruction in the same way as Fourier feature mappings help coordinate-based MLPs overcome spectral bias. Moreover, we theoretically analyze Matérn kernels' connection to SIREN networks as well as their relation to previously employed arc-cosine kernels. Finally, based on recently introduced Neural Kernel Fields, we present data-dependent Matérn kernels and conclude that especially the Laplace kernel (being part of the Matérn family) is extremely competitive, performing almost on par with state-of-the-art methods in the noise-free case while having a more than five times shorter training time.

1983. Brain-inspired L_p-Convolution benefits large kernels and aligns better with visual cortex

链接：https://iclr.cc/virtual/2025/poster/31259 abstract： Convolutional Neural Networks (CNNs) have profoundly influenced the field of computer vision, drawing significant inspiration from the visual processing mechanisms inherent in the brain. Despite sharing fundamental structural and representational similarities with the biological visual system, differences in local connectivity patterns within CNNs open up an interesting area to explore. In this work, we explore whether integrating biologically observed receptive fields (RFs) can enhance model performance and foster alignment with brain representations. We introduce a novel methodology, termed L_p-convolution, which employs the multivariate L_p-generalized normal distribution as an adaptable L_p-masks, to reconcile disparities between artificial and biological RFs. L_p-masks finds the optimal RFs through task-dependent adaptation of conformation such as distortion, scale, and rotation. This allows L_p-convolution to excel in tasks that require flexible RF shapes, including not only square-shaped regular RFs but also horizontal and vertical ones. Furthermore, we demonstrate that L_p-convolution with biological RFs significantly enhances the performance of large kernel CNNs possibly by introducing structured sparsity inspired by L_p-generalized normal distribution in convolution. Lastly, we present that neural representations of CNNs align more closely with the visual cortex when -convolution is close to biological RFs.

1984. In Search of the Engram in LLMs: A Neuroscience Perspective on the Memory Functions in AI Models

链接：https://iclr.cc/virtual/2025/poster/31329 abstract： Large Language Models (LLMs) are enhancing our daily lives but also pose risks like spreading misinformation and violating privacy, highlighting the importance of understanding how they process and store information. This blogpost offers a fresh look into a neuroscience-inspired perspective of LLM's memory functions, based on the concept of engrams-the physical substrate of memory in living organism. We discuss a synergy between AI research and neuroscience, as both fields cover complexities of intelligent systems.

1985. GLoRa: A Benchmark to Evaluate the Ability to Learn Long-Range Dependencies in Graphs

链接：https://iclr.cc/virtual/2025/poster/31120 abstract： Learning on graphs is one of the most active research topics in

machine learning (ML). Among the key challenges in this field, effectively learning long-range dependencies in graphs has been particularly difficult. It has been observed that, in practice, the performance of many ML approaches, including various types of graph neural networks (GNNs), degrades significantly when the learning task involves long-range dependencies—that is, when the answer is determined by the presence of a certain path of significant length in the graph. This issue has been attributed to several phenomena, including over-smoothing, over-squashing, and vanishing gradient. A number of solutions have been proposed to mitigate these causes. However, evaluation of these solutions is currently challenging because existing benchmarks do not effectively test systems for their ability to learn tasks based on long-range dependencies in a transparent manner. In this paper, we introduce GLoRa, a synthetic benchmark that allows testing of systems for this ability in a systematic way. We then evaluate state-of-the-art systems using GLoRa and conclude that none of them can confidently claim to learn long-range dependencies well. We also observe that this weak performance cannot be attributed to any of the three causes, highlighting the need for further investigation.

1986. Progress or Regress? Self-Improvement Reversal in Post-training

链接：https://iclr.cc/virtual/2025/poster/29657 abstract： Self-improvement through post-training methods such as iterative preference learning has been acclaimed for enhancing the problem-solving capabilities (e.g., mathematical reasoning) of Large Language Models (LLMs) without human intervention. However, as our exploration deepens, it is crucial to critically assess whether these enhancements indeed signify comprehensive progress or if they could lead to unintended regressions. Through rigorous experimentation and analysis across diverse problem-solving tasks, we uncover nuances in the self-improvement trajectories of LLMs. Our study introduces the concept of \emph{self-improvement reversal}, where models showing improved overall accuracy metrics might paradoxically exhibit declines in broader, essential capabilities. We propose a comprehensive evaluative framework to scrutinize the underlying mechanisms and outcomes of post-training self-improvement, aiming to discern between superficial metric improvements and genuine enhancements in model functionality. The findings emphasize the complexity of technological advancements in LLMs, underscoring the need for a nuanced understanding of the \textit{progress or regress} dichotomy in their development.

1987. MAESTRO: Masked Encoding Set Transformer with Self-Distillation

链接：https://iclr.cc/virtual/2025/poster/30353 abstract： The interrogation of cellular states and interactions in immunology research is an ever-evolving task, requiring adaptation to the current levels of high dimensionality. Cytometry enables high-dimensional profiling of immune cells, but its analysis is hindered by the complexity and variability of the data. We present MAESTRO, a self-supervised set representation learning model that generates vector representations of set-structured data, which we apply to learn immune profiles from cytometry data. Unlike previous studies only learn cell-level representations, whereas MAESTRO uses all of a sample's cells to learn a set representation. MAESTRO leverages specialized attention mechanisms to handle sets of variable number of cells and ensure permutation invariance, coupled with an online tokenizer by self-distillation framework. We benchmarked our model against existing cytometry approaches and other existing machine learning methods that have never been applied in cytometry. Our model outperforms existing approaches in retrieving cell-type proportions and capturing clinically relevant features for downstream tasks such as disease diagnosis and immune cell profiling.

1988. dEBORA: Efficient Bilevel Optimization-based low-Rank Adaptation

链接：https://iclr.cc/virtual/2025/poster/30949 abstract： Low-rank adaptation methods are a popular approach for parameter-efficient fine-tuning of large-scale neural networks. However, selecting the optimal rank for each layer remains a challenging problem that significantly affects both performance and efficiency. In this paper, we introduce a novel bilevel optimization strategy that simultaneously trains both matrix and tensor low-rank adapters, dynamically selecting the optimal rank for each layer. Our method avoids the use of implicit differentiation in the computation of the hypergradient, and integrates a stochastic away-step variant of the Frank-Wolfe algorithm, eliminating the need for projection and providing identifiability guarantees of the optimal rank structure. This results in a highly efficient and cost-effective training scheme that adaptively allocates the parameter budget across the network layers. On top of a detailed theoretical analysis of the method, we provide different numerical experiments showcasing its effectiveness.

1989. DataEnvGym: Data Generation Agents in Teacher Environments with Student Feedback

链接：https://iclr.cc/virtual/2025/poster/31276 abstract： The process of creating training data to teach models is currently driven by humans, who manually analyze model weaknesses and plan how to create data that improves a student model. Recent approaches using large language models (LLMs) as annotators reduce human annotation effort, but still require humans to interpret feedback from evaluations and control the LLM to produce data the student needs. Automating this labor-intensive process by creating autonomous data generation agents – or teachers – is desirable, but requires environments that can simulate the feedback-driven, iterative, closed loop of data creation. To enable rapid and scalable testing for such agents and their modules, we introduce DataEnvGym, a testbed of teacher environments for data generation agents. DataEnvGym frames data generation as a sequential decision-making task, involving an agent consisting of a data generation policy (which generates a plan for creating training data) and a data generation engine (which transforms the plan into data), inside an environment that provides feedback from a student. The agent’s end goal is to improve student model performance. Students are iteratively trained and evaluated on generated data, with their feedback (in the form of errors or weak skills) being reported to the agent after each iteration. As a general-purpose testbed, DataEnvGym includes multiple instantiations of teacher

environments across three levels of structure in the state representation and action space, with varying levels of scaffolding support. More structured environments are based on automatically-inferred skills and offer a higher degree of interpretability and control over the curriculum. We support developing and testing data generation agents in four diverse tasks covering text, images, and actions (mathematics, programming, visual question answering, and tool-use) and test multiple student and teacher models. We find that example agents in our teaching environments can iteratively improve students across diverse tasks and settings. Moreover, we show that environments can teach different skill levels and can be used to test variants of key modules, pointing to directions of future work in improving data generation agents, engines, and feedback mechanisms. Project page: https://DataEnvGym.github.io.

1990. Endless Jailbreaks with Bijection Learning

链接：https://iclr.cc/virtual/2025/poster/27786 abstract： Despite extensive safety measures, LLMs are vulnerable to adversarial inputs, or jailbreaks, which can elicit unsafe behaviors. In this work, we introduce bijection learning, a powerful attack algorithm which automatically fuzzes LLMs for safety vulnerabilities using randomly-generated encodings whose complexity can be tightly controlled. We leverage in-context learning to teach models bijective encodings, pass encoded queries to the model to bypass built-in safety mechanisms, and finally decode responses back into English. Our attack is extremely effective on a wide range of frontier language models. By controlling complexity parameters such as number of key-value mappings in the encodings, we find a close relationship between the capability level of the attacked LLM and the average complexity of the most effective bijection attacks. Our work highlights that new vulnerabilities in frontier models can emerge with scale: more capable models are more severely jailbroken by bijection attacks.

1991. Multi-session, multi-task neural decoding from distinct cell-types and brain regions

链接：https://iclr.cc/virtual/2025/poster/30143 abstract： Recent work has shown that scale is important for improved brain decoding, with more data leading to greater decoding accuracy. However, large-scale decoding across many different datasets is challenging because neural circuits are heterogeneous---each brain region contains a unique mix of cellular sub-types, and the responses to different stimuli are diverse across regions and sub-types. It is unknown whether it is possible to pre-train and transfer brain decoding models between distinct tasks, cellular sub-types, and brain regions. To address these questions, we developed a multi-task transformer architecture and trained it on the entirety of the Allen Institute's Brain Observatory dataset. This dataset contains responses from over 100,000 neurons in 6 areas of the brains of mice, observed with two-photon calcium imaging, recorded while the mice observed different types of visual stimuli. Our results demonstrate that transfer is indeed possible -combining data from different sources is beneficial for a number of downstream decoding tasks. As well, we can transfer the model between regions and sub-types, demonstrating that there is in fact common information in diverse circuits that can be extracted by an appropriately designed model. Interestingly, we found that the model's latent representations showed clear distinctions between different brain regions and cellular sub-types, even though it was never given any information about these distinctions. Altogether, our work demonstrates that training a large-scale neural decoding model on diverse data is possible, and this provides a means of studying the differences and similarities between heterogeneous neural circuits.

1992. Towards Multiple Character Image Animation Through Enhancing Implicit Decoupling

链接：https://iclr.cc/virtual/2025/poster/29146 abstract： Controllable character image animation has a wide range of applications. Although existing studies have consistently improved performance, challenges persist in the field of character image animation, particularly concerning stability in complex backgrounds and tasks involving multiple characters. To address these challenges, we propose a novel multi-condition guided framework for character image animation, employing several well-designed input modules to enhance the implicit decoupling capability of the model. First, the optical flow guider calculates the background optical flow map as guidance information, which enables the model to implicitly learn to decouple the background motion into background constants and background momentum during training, and generate a stable background by setting zero background momentum during inference. Second, the depth order guider calculates the order map of the characters, which transforms the depth information into the positional information of multiple characters. This facilitates the implicit learning of decoupling different characters, especially in accurately separating the occluded body parts of multiple characters. Third, the reference pose map is input to enhance the ability to decouple character texture and pose information in the reference image. Furthermore, to fill the gap of fair evaluation of multi-character image animation, we propose a new benchmark comprising about 4,000 frames. Extensive qualitative and quantitative evaluations demonstrate that our method excels in generating high-quality character animations, especially in scenarios of complex backgrounds and multiple characters.

1993. Offline Model-Based Optimization by Learning to Rank

链接：https://iclr.cc/virtual/2025/poster/28117 abstract： Offline model-based optimization (MBO) aims to identify a design that maximizes a black-box function using only a fixed, pre-collected dataset of designs and their corresponding scores. This problem has garnered significant attention from both scientific and industrial domains. A common approach in offline MBO is to train a regression-based surrogate model by minimizing mean squared error (MSE) and then find the best design within this surrogate model by different optimizers (e.g., gradient ascent). However, a critical challenge is the risk of out-of-distribution errors, i.e., the surrogate model may typically overestimate the scores and mislead the optimizers into suboptimal regions. Prior

works have attempted to address this issue in various ways, such as using regularization techniques and ensemble learning to enhance the robustness of the model, but it still remains. In this paper, we argue that regression models trained with MSE are not well-aligned with the primary goal of offline MBO, which is to \textit{select} promising designs rather than to predict their scores precisely. Notably, if a surrogate model can maintain the order of candidate designs based on their relative score relationships, it can produce the best designs even without precise predictions. To validate it, we conduct experiments to compare the relationship between the quality of the final designs and MSE, finding that the correlation is really very weak. In contrast, a metric that measures order-maintaining quality shows a significantly stronger correlation. Based on this observation, we propose learning a ranking-based model that leverages learning to rank techniques to prioritize promising designs based on their relative scores. We show that the generalization error on ranking loss can be well bounded. Empirical results across diverse tasks demonstrate the superior performance of our proposed ranking-based method than twenty existing methods. Our implementation is available at \url{https://github.com/lamda-bbo/Offline-RaM}.

1994. Open-CK: A Large Multi-Physics Fields Coupling benchmarks in Combustion Kinetics

链接：https://iclr.cc/virtual/2025/poster/30654 abstract： In this paper, we use the Fire Dynamics Simulator (FDS) combined with the {\fontfamily{lmtt}\selectfont \textit{supercomputer}} support to create a \textbf{C}ombustion \textbf{K}inetics (CK) dataset for machine learning and scientific research. This dataset captures the development of fires in industrial parks with high-precision Computational Fluid Dynamics (CFD) simulations. It includes various physical fields such as temperature and pressure, and covers multiple environmental combinations for exploring \underline{multi-physics} field coupling phenomena. Additionally, we evaluate several advanced machine learning architectures across our {\fontfamily{lmtt}\selectfont {Open-CK}} benchmark using a substantial computational setup of 64 NVIDIA A100 GPUs: \ding{182} vision backbone; \ding{183} spatio-temporal predictive models; \ding{184} operator learning frameworks. These architectures uniquely excel at handling complex physical field data. We also introduce three benchmarks to demonstrate their potential in enhancing the exploration of downstream tasks: (a) capturing continuous changes in combustion kinetics; (b) a neural partial differential equation solver for learning temperature fields and turbulence; (c) reconstruction of sparse physical observations. The Open-CK dataset and benchmarks aim to advance research in combustion kinetics driven by machine learning, providing a reliable baseline for developing and comparing cutting-edge technologies and models. We hope to further promote the application of deep learning in earth sciences. Our project is available at \url{https://github.com/whscience/Open-CK}.

1995. Self-Improvement in Language Models: The Sharpening Mechanism

链接：https://iclr.cc/virtual/2025/poster/29372 abstract： Recent work in language modeling has raised the possibility of “self-improvement,” where an LLM evaluates and refines its own generations to achieve higher performance without external feedback. It is impossible for this self-improvement to create information that is not already in the model, so why should we expect that this will lead to improved capabilities? We offer a new theoretical perspective on the capabilities of self-improvement through a lens we refer to as “sharpening.” Motivated by the observation that language models are often better at verifying response quality than they are at generating correct responses, we formalize self-improvement as using the model itself as a verifier during post-training in order to ‘sharpen’ the model to one placing large mass on high-quality sequences, thereby amortizing the expensive inference-time computation of generating good sequences. We begin by introducing a new statistical framework for sharpening in which the learner has sample access to a pre-trained base policy. Then, we analyze two natural families of self improvement algorithms based on SFT and RLHF. We find that (i) the SFT-based approach is minimax optimal whenever the initial model has sufficient coverage, but (ii) the RLHF-based approach can improve over SFT-based self-improvement by leveraging online exploration, bypassing the need for coverage. We view these findings as a starting point toward a foundational understanding that can guide the design and evaluation of self-improvement algorithms.

1996. RecDreamer: Consistent Text-to-3D Generation via Uniform Score Distillation

链接：https://iclr.cc/virtual/2025/poster/29139 abstract： Current text-to-3D generation methods based on score distillation often suffer from geometric inconsistencies, leading to repeated patterns across different poses of 3D assets. This issue, known as the Multi-Face Janus problem, arises because existing methods struggle to maintain consistency across varying poses and are biased toward a canonical pose. While recent work has improved pose control and approximation, these efforts are still limited by this inherent bias, which skews the guidance during generation.To address this, we propose a solution called RecDreamer, which reshapes the underlying data distribution to achieve more consistent pose representation. The core idea behind our method is to rectify the prior distribution, ensuring that pose variation is uniformly distributed rather than biased toward a canonical form. By modifying the prescribed distribution through an auxiliary function, we can reconstruct the density of the distribution to ensure compliance with specific marginal constraints. In particular, we ensure that the marginal distribution of poses follows a uniform distribution, thereby eliminating the biases introduced by the prior knowledge.We incorporate this rectified data distribution into existing score distillation algorithms, a process we refer to as uniform score distillation. To efficiently compute the posterior distribution required for the auxiliary function, RecDreamer introduces a training-free classifier that estimates pose categories in a plug-and-play manner. Additionally, we utilize various approximation techniques for noisy states, significantly improving system performance.Our experimental results demonstrate that RecDreamer effectively mitigates the Multi-Face Janus problem, leading to more consistent 3D asset generation across different poses.

1997. Nonasymptotic Analysis of Stochastic Gradient Descent with the Richardson–Romberg Extrapolation

链接：https://iclr.cc/virtual/2025/poster/29815 abstract： We address the problem of solving strongly convex and smooth minimization problems using stochastic gradient descent (SGD) algorithm with a constant step size. Previous works suggested to combine the Polyak-Ruppert averaging procedure with the Richardson-Romberg extrapolation to reduce the asymptotic bias of SGD at the expense of a mild increase of the variance. We significantly extend previous results by providing an expansion of the mean-squared error of the resulting estimator with respect to the number of iterations n. We show that the root mean-squared error can be decomposed into the sum of two terms: a leading one of order $\mathcal{O}(n^{-1/2})$ with explicit dependence on a minimax-optimal asymptotic covariance matrix, and a second-order term of order $\mathcal{O}(n^{-3/4})$, where the power $3/4$ is best known. We also extend this result to the higher-order moment bounds. Our analysis relies on the properties of the SGD iterates viewed as a time-homogeneous Markov chain. In particular, we establish that this chain is geometrically ergodic with respect to a suitably defined weighted Wasserstein semimetric.

1998. CameraCtrl: Enabling Camera Control for Video Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29236 abstract： Controllability plays a crucial role in video generation, as it allows users to create and edit content more precisely. Existing models, however, lack control of camera pose that serves as a cinematic language to express deeper narrative nuances. To alleviate this issue, we introduce \method, enabling accurate camera pose control for video diffusion models. Our approach explores effective camera trajectory parameterization along with a plug-and-play camera pose control module that is trained on top of a video diffusion model, leaving other modules of the base model untouched. Moreover, a comprehensive study on the effect of various training datasets is conducted, suggesting that videos with diverse camera distributions and similar appearance to the base model indeed enhance controllability and generalization. Experimental results demonstrate the effectiveness of \method in achieving precise camera control with different video generation models, marking a step forward in the pursuit of dynamic and customized video storytelling from textual and camera pose inputs.

1999. RoboCat: A Self-Improving Generalist Agent for Robotic Manipulation

链接：https://iclr.cc/virtual/2025/poster/31507 abstract： The ability to leverage heterogeneous robotic experience from different robots and tasks to quickly master novel skills and embodiments has the potential to transform robot learning. Inspired by recent advances in foundation models for vision and language, we propose a multi-embodiment, multi-task generalist agent for robotic manipulation. This agent, named RoboCat, is a visual goal-conditioned decision transformer capable of consuming action-labelled visual experience. This data spans a large repertoire of motor control skills from simulated and real robotic arms with varying sets of observations and actions. With RoboCat, we demonstrate the ability to generalise to new tasks and robots, both zero-shot as well as through adaptation using only 100–1000 examples for the target task. We also show how a trained model itself can be used to generate data for subsequent training iterations, thus providing a basic building block for an autonomous improvement loop. We investigate the agent's capabilities, with large-scale evaluations both in simulation and on three different real robot embodiments. We find that as we grow and diversify its training data, RoboCat not only shows signs of cross-task transfer, but also becomes more efficient at adapting to new tasks.

2000. Feature-Based Online Bilateral Trade

链接：https://iclr.cc/virtual/2025/poster/27764 abstract： Bilateral trade models the problem of facilitating trades between a seller and a buyer having private valuations for the item being sold. In the online version of the problem, the learner faces a new seller and buyer at each time step, and has to post a price for each of the two parties without any knowledge of their valuations. We consider a scenario where, at each time step, before posting prices the learner observes a context vector containing information about the features of the item for sale. The valuations of both the seller and the buyer follow an unknown linear function of the context. In this setting, the learner could leverage previous transactions in an attempt to estimate private valuations. We characterize the regret regimes of different settings, taking as a baseline the best context-dependent prices in hindsight. First, in the setting in which the learner has two-bit feedback and strong budget balance constraints, we propose an algorithm with $O(\log T)$ regret. Then, we study the same set-up with noisy valuations, providing a tight $\widetilde O(T^{2/3})$ regret upper bound. Finally, we show that loosening budget balance constraints allows the learner to operate under more restrictive feedback. Specifically, we show how to address the one-bit, global budget balance setting through a reduction from the two-bit, strong budget balance setup. This established a fundamental trade-off between the quality of the feedback and the strictness of the budget constraints.