2001. DiffSplat: Repurposing Image Diffusion Models for Scalable Gaussian Splat Generation

链接：https://iclr.cc/virtual/2025/poster/28918 abstract： Recent advancements in 3D content generation from text or a single image struggle with limited high-quality 3D datasets and inconsistency from 2D multi-view generation. We introduce DiffSplat, a novel 3D generative framework that natively generates 3D Gaussian splats by taming large-scale text-to-image diffusion models. It differs from previous 3D generative models by effectively utilizing web-scale 2D priors while maintaining 3D consistency in a unified model. To bootstrap the training, a lightweight reconstruction model is proposed to instantly produce multi-view Gaussian splat grids for scalable dataset curation. In conjunction with the regular diffusion loss on these grids, a 3D rendering loss is introduced to facilitate 3D coherence across arbitrary views. The compatibility with image diffusion models enables seamless adaptions of numerous techniques for image generation to the 3D realm. Extensive experiments reveal the superiority of DiffSplat in text- and image-conditioned generation tasks and downstream applications. Thorough ablation studies validate the efficacy of each critical design choice and provide insights into the underlying mechanism.

2002. Layout-your-3D: Controllable and Precise 3D Generation with 2D Blueprint

链接：https://iclr.cc/virtual/2025/poster/28437 abstract： We present Layout-Your-3D, a framework that allows controllable and compositional 3D generation from text prompts. Existing text-to-3D methods often struggle to generate assets with plausible object interactions or require tedious optimization processes. To address these challenges, our approach leverages 2D layouts as a blueprint to facilitate precise and plausible control over 3D generation. Starting with a 2D layout provided by a user or generated from a text description, we first create a coarse 3D scene using a carefully designed initialization process based on efficient reconstruction models. To enforce coherent global 3D layouts and enhance the quality of instance appearances, we propose a collision-aware layout optimization process followed by instance-wise refinement. Experimental results demonstrate that Layout-Your-3D yields more reasonable and visually appealing compositional 3D assets while significantly reducing the time required for each prompt. Additionally, Layout-Your-3D can be easily applicable to downstream tasks, such as 3D editing and object insertion.

2003. No Pose, No Problem: Surprisingly Simple 3D Gaussian Splats from Sparse Unposed Images

链接：https://iclr.cc/virtual/2025/poster/29779 abstract： We introduce NoPoSplat, a feed-forward model capable of reconstructing 3D scenes parameterized by 3D Gaussians from unposed sparse multi-view images. Our model, trained exclusively with photometric loss, achieves real-time 3D Gaussian reconstruction during inference. To eliminate the need for accurate pose input during reconstruction, we anchor one input view's local camera coordinates as the canonical space and train the network to predict Gaussian primitives for all views within this space. This approach obviates the need to transform Gaussian primitives from local coordinates into a global coordinate system, thus avoiding errors associated with per-frame Gaussians and pose estimation. To resolve scale ambiguity, we design and compare various intrinsic embedding methods, ultimately opting to convert camera intrinsics into a token embedding and concatenate it with image tokens as input to the model, enabling accurate scene scale prediction. We utilize the reconstructed 3D Gaussians for novel view synthesis and pose estimation tasks and propose a two-stage coarse-to-fine pipeline for accurate pose estimation. Experimental results demonstrate that our pose-free approach can achieve superior novel view synthesis quality compared to pose-required methods, particularly in scenarios with limited input image overlap. For pose estimation, our method, trained without ground truth depth or explicit matching loss, significantly outperforms the state-of-the-art methods with substantial improvements. This work makes significant advances in pose-free generalizable 3D reconstruction and demonstrates its applicability to real-world scenarios. Code and trained models are available at https://noposplat.github.io/.

2004. Comparing Targeting Strategies for Maximizing Social Welfare with Limited Resources

链接：https://iclr.cc/virtual/2025/poster/31240 abstract： Machine learning is increasingly used to select which individuals receive limited-resource interventions in domains such as human services, education, development, and more. However, it is often not apparent what the right quantity is for models to predict. In particular, policymakers rarely have access to data from a randomized controlled trial (RCT) that would enable accurate estimates of treatment effects -- which individuals would benefit more from the intervention. Observational data is more likely to be available, creating a substantial risk of bias in treatment effect estimates. Practitioners instead commonly use a technique termed "risk-based targeting" where the model is just used to predict each individual's status quo outcome (an easier, non-causal task). Those with higher predicted risk are offered treatment. There is currently almost no empirical evidence to inform which choices lead to the most effect machine learning-informed targeting strategies in social domains. In this work, we use data from 5 real-world RCTs in a variety of domains to empirically assess such choices. We find that when treatment effects can be estimated reliably (which we simulate by using direct outcome observations), treatment effect based targeting substantially outperforms risk-based targeting, even when treatment effect estimates are biased. Moreover, these results hold even when the policymaker has strong normative preferences for assisting higher-risk individuals. However, when treatment effects must be predicted from features alone (as is always the case in practice), performance can degrade significantly due to limited data making it difficult to learn accurate mappings from features to treatment effects. Our results suggest treatment effect targeting has significant potential benefits, but realizing these benefits requires careful attention to model training and validation.

2005. Dynamic Sparse Training versus Dense Training: The Unexpected Winner in Image Corruption Robustness

链接：https://iclr.cc/virtual/2025/poster/28977 abstract： It is generally perceived that Dynamic Sparse Training opens the door to a new era of scalability and efficiency for artificial neural networks at, perhaps, some costs in accuracy performance for the classification task. At the same time, Dense Training is widely accepted as being the "de facto" approach to train artificial neural networks if one would like to maximize their robustness against image corruption. In this paper, we question this general practice. Consequently, \textit{we claim that}, contrary to what is commonly thought, the Dynamic Sparse Training methods can consistently outperform Dense Training in terms of robustness accuracy, particularly if the efficiency aspect is not considered as a main objective (i.e., sparsity levels between 10\% and up to 50\%), without adding (or even reducing) resource cost. We validate our claim on two types of data, images and videos, using several traditional and modern deep learning architectures for computer vision and three widely studied Dynamic Sparse Training algorithms. Our findings reveal a new yet-unknown benefit of Dynamic Sparse Training and open new possibilities in improving deep learning robustness beyond the current state of the art.

2006. MIA-DPO: Multi-Image Augmented Direct Preference Optimization For Large Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/28892 abstract： Visual preference alignment involves training Large Vision-Language Models (LVLMs) to predict human preferences between visual inputs. This is typically achieved by using labeled datasets of chosen/rejected pairs and employing optimization algorithms like direct preference optimization (DPO).Existing visual alignment methods, primarily designed for single-image scenarios, struggle to effectively handle the complexity of multi-image tasks due to the scarcity of diverse training data and the high cost of annotating chosen/rejected pairs.We present Multi-Image Augmented Direct Preference Optimization (MIA-DPO), a visual preference alignment approach that effectively handles multi-image inputs.MIA-DPO mitigates the scarcity of diverse multi-image training data by extending single-image data with unrelated images arranged in grid collages or pic-in-pic formats, significantly reducing the costs associated with multi-image data annotations.Our observation reveals that attention values of LVLMs vary considerably across different images. We use attention values to identify and filter out rejected responses the model may have mistakenly focused on.Our attention-aware selection for constructing the chosen/rejected pairs without relying on (i) human annotation, (ii) extra data, and (iii) external models or APIs.MIA-DPO is compatible with various architectures and outperforms existing methods on five multi-image benchmarks, achieving an average performance boost of 3.0% on LLaVA-v1.5 and 4.3% on the recent InternLM-XC2.5.Moreover, MIA-DPO has a minimal effect on the model's ability to understand single images.

2007. MotionClone: Training-Free Motion Cloning for Controllable Video Generation

链接：https://iclr.cc/virtual/2025/poster/29159 abstract： Motion-based controllable video generation offers the potential for creating captivating visual content. Existing methods typically necessitate model training to encode particular motion cues or incorporate fine-tuning to inject certain motion patterns, resulting in limited flexibility and generalization. In this work, we propose MotionClone, a training-free framework that enables motion cloning from reference videos to versatile motion-controlled video generation, including text-to-video and image-to-video. Based on the observation that the dominant components in temporal-attention maps drive motion synthesis, while the rest mainly capture noisy or very subtle motions, MotionClone utilizes sparse temporal attention weights as motion representations for motion guidance, facilitating diverse motion transfer across varying scenarios. Meanwhile, MotionClone allows for the direct extraction of motion representation through a single denoising step, bypassing the cumbersome inversion processes and thus promoting both efficiency and flexibility. Extensive experiments demonstrate that MotionClone exhibits proficiency in both global camera motion and local object motion, with notable superiority in terms of motion fidelity, textual alignment, and temporal consistency.

2008. Systematic Relational Reasoning With Epistemic Graph Neural Networks

链接：https://iclr.cc/virtual/2025/poster/28252 abstract： Developing models that can learn to reason is a notoriously challenging problem. We focus on reasoning in relational domains, where the use of Graph Neural Networks (GNNs) seems like a natural choice. However, previous work has shown that regular GNNs lack the ability to systematically generalize from training examples on test graphs requiring longer inference chains, which fundamentally limits their reasoning abilities. A common solution relies on neuro-symbolic methods that systematically reason by learning rules, but their scalability is often limited and they tend to make unrealistically strong assumptions, e.g.\ that the answer can always be inferred from a single relational path. We propose the Epistemic GNN (EpiGNN), a novel parameter-efficient and scalable GNN architecture with an epistemic inductive bias for systematic reasoning. Node embeddings in EpiGNNs are treated as epistemic states, and message passing is implemented accordingly. We show that EpiGNNs achieve state-of-the-art results on link prediction tasks that require systematic reasoning. Furthermore, for inductive knowledge graph completion, EpiGNNs rival the performance of state-of-the-art specialized approaches. Finally, we introduce two new benchmarks that go beyond standard relational reasoning by requiring the aggregation of information from multiple paths. Here, existing neuro-symbolic approaches fail, yet EpiGNNs learn to reason accurately. Code and datasets are available at https://github.com/erg0dic/gnn-sg.

2009. CapeX: Category-Agnostic Pose Estimation from Textual Point Explanation

链接：https://iclr.cc/virtual/2025/poster/28113 abstract： Conventional 2D pose estimation models are constrained by their design to specific object categories. This limits their applicability to predefined objects. To overcome these limitations, category-agnostic pose estimation (CAPE) emerged as a solution. CAPE aims to facilitate keypoint localization for diverse object categories using a unified model, which can generalize from minimal annotated support images.Recent CAPE works have produced object poses based on arbitrary keypoint definitions annotated on a user-provided support image. Our work departs from conventional CAPE methods, which require a support image, by adopting a text-based approach instead of the support image. Specifically, we use a pose-graph, where nodes represent keypoints that are described with text. This representation takes advantage of the abstraction of text descriptions and the structure imposed by the graph.Our approach effectively breaks symmetry, preserves structure, and improves occlusion handling.We validate our novel approach using the MP-100 benchmark, a comprehensive dataset covering over 100 categories and 18,000 images. MP-100 is structured so that the evaluation categories are unseen during training, making it especially suited for CAPE. Under a 1-shot setting, our solution achieves a notable performance boost of 1.26\%, establishing a new state-of-the-art for CAPE. Additionally, we enhance the dataset by providing text description annotations for both training and testing. We also include alternative text annotations specifically for testing the model's ability to generalize across different textual descriptions, further increasing its value for future research. Our code and dataset are publicly available at https://github.com/matanr/capex.

2010. System 1.x: Learning to Balance Fast and Slow Planning with Language Models

链接：https://iclr.cc/virtual/2025/poster/27660 abstract： Language models can be used to solve long-horizon planning problems in two distinct modes. In a fast 'System-1' mode, models directly generate plans without any explicit search or backtracking, and in a slow 'System-2' mode, they plan step-by-step by explicitly searching over possible actions. System-2 planning, while typically more effective, is also computationally more expensive and often infeasible for long plans or large action spaces. Moreover, isolated System-1 or System-2 planning ignores the user's end goals and constraints (e.g., token budget), failing to provide ways for the user to control the model's behavior. To this end, we propose the System-1.x Planner, a framework for controllable planning with language models that is capable of generating hybrid plans and balancing between the two planning modes based on the difficulty of the problem at hand. System-1.x consists of (i) a controller, (ii) a System-1 Planner, and (iii) a System-2 Planner. Based on a user-specified hybridization factor x governing the degree to which the system uses System-1 vs. System-2, the controller decomposes a planning problem into subgoals, and classifies them as easy or hard to be solved by either System-1 or System-2, respectively. We fine-tune all three components on top of a single base LLM, requiring only search traces as supervision. Experiments with two diverse planning tasks -- Maze Navigation and Blocksworld -- show that our System-1.x Planner outperforms a System-1 Planner, a System-2 Planner trained to approximate A *search, and also a symbolic planner (A* search), given a state exploration budget. We also demonstrate the following key properties of our planner: (1) controllability: by adjusting the hybridization factor x (e.g., System-1.75 vs. System-1.5) we can perform more (or less) search, improving performance, (2) flexibility: by building a neuro-symbolic variant composed of a neural System-1 planner and a symbolic System-2 planner, we can take advantage of existing symbolic methods, and (3) generalizability: by learning from different search algorithms (BFS, DFS, A*), we show that our method is robust to the choice of search algorithm used for training.

2011. LLM-wrapper: Black-Box Semantic-Aware Adaptation of Vision-Language Models for Referring Expression Comprehension

链接：https://iclr.cc/virtual/2025/poster/29739 abstract： Vision Language Models (VLMs) have demonstrated remarkable capabilities in various open-vocabulary tasks, yet their zero-shot performance lags behind task-specific fine-tuned models, particularly in complex tasks like Referring Expression Comprehension (REC). Fine-tuning usually requires ‘white-box’ access to the model’s architecture and weights, which is not always feasible due to proprietary or privacy concerns. In this work, we propose LLM-wrapper, a method for ‘black-box’ adaptation of VLMs for the REC task using Large Language Models (LLMs). LLM-wrapper capitalizes on the reasoning abilities of LLMs, improved with a light fine-tuning, to select the most relevant bounding box to match the referring expression, from candidates generated by a zero-shot black-box VLM. Our approach offers several advantages: it enables the adaptation of closed-source models without needing access to their internal workings, it is versatile and works with any VLM, transfers to new VLMs, and it allows for the adaptation of an ensemble of VLMs. We evaluate LLM-wrapper on multiple datasets using different VLMs and LLMs, demonstrating significant performance improvements and highlighting the versatility of our method. While LLM-wrapper is not meant to directly compete with standard white-box fine-tuning, it offers a practical and effective alternative for black-box VLM adaptation. The code will be open-sourced.

2012. Understanding the Generalization of In-Context Learning in Transformers: An Empirical Study

链接：https://iclr.cc/virtual/2025/poster/27730 abstract： Large language models (LLMs) like GPT-4 and LLaMA-3 utilize the powerful in-context learning (ICL) capability of Transformer architecture to learn on the fly from limited examples. While ICL underpins many LLM applications, its full potential remains hindered by a limited understanding of its generalization boundaries and vulnerabilities. We present a systematic investigation of transformers' generalization capability with ICL relative to training data coverage by defining a task-centric framework along three dimensions: inter-problem, intra-problem, and intra-task generalization. Through extensive simulation and real-world experiments, encompassing tasks such as function fitting, API calling, and translation, we find that transformers lack inter-problem generalization with ICL, but excel in intra-task and intra-problem generalization. When the training data includes a greater variety of mixed tasks, it significantly enhances the generalization ability of ICL on unseen tasks and even on known simple tasks. This guides us in designing training data to maximize the diversity of tasks covered and to combine different tasks whenever possible, rather than solely focusing on the target task for testing.

2013. Halton Scheduler for Masked Generative Image Transformer

链接：https://iclr.cc/virtual/2025/poster/29658 abstract： Masked Generative Image Transformers (MaskGIT) have emerged as a scalableand efficient image generation framework, able to deliver high-quality visuals withlow inference costs. However, MaskGIT’s token unmasking scheduler, an essentialcomponent of the framework, has not received the attention it deserves. We analyzethe sampling objective in MaskGIT, based on the mutual information betweentokens, and elucidate its shortcomings. We then propose a new sampling strategybased on our Halton scheduler instead of the original Confidence scheduler. Moreprecisely, our method selects the token’s position according to a quasi-random,low-discrepancy Halton sequence. Intuitively, that method spreads the tokensspatially, progressively covering the image uniformly at each step. Our analysisshows that it allows reducing non-recoverable sampling errors, leading to simplerhyper-parameters tuning and better quality images. Our scheduler does not requireretraining or noise injection and may serve as a simple drop-in replacement forthe original sampling strategy. Evaluation of both class-to-image synthesis onImageNet and text-to-image generation on the COCO dataset demonstrates that theHalton scheduler outperforms the Confidence scheduler quantitatively by reducingthe FID and qualitatively by generating more diverse and more detailed images.Our code is at https://github.com/valeoai/Halton-MaskGIT.

2014. Continuity-Preserving Convolutional Autoencoders for Learning Continuous Latent Dynamical Models from Images

链接：https://iclr.cc/virtual/2025/poster/29910 abstract： Continuous dynamical systems are cornerstones of many scientific and engineering disciplines.While machine learning offers powerful tools to model these systems from trajectory data, challenges arise when these trajectories are captured as images, resulting in pixel-level observations that are discrete in nature.Consequently, a naive application of a convolutional autoencoder can result in latent coordinates that are discontinuous in time.To resolve this, we propose continuity-preserving convolutional autoencoders (CpAEs) to learn continuous latent states and their corresponding continuous latent dynamical models from discrete image frames. We present a mathematical formulation for learning dynamics from image frames, which illustrates issues with previous approaches and motivates our methodology based on promoting the continuity of convolution filters, thereby preserving the continuity of the latent states.This approach enables CpAEs to produce latent states that evolve continuously with the underlying dynamics, leading to more accurate latent dynamical models.Extensive experiments across various scenarios demonstrate the effectiveness of CpAEs.

2015. σ-zero: Gradient-based Optimization of ℓ_0-norm Adversarial Examples

链接：https://iclr.cc/virtual/2025/poster/30114 abstract： Evaluating the adversarial robustness of deep networks to gradient-based attacks is challenging.While most attacks consider ℓ_2- and ℓ_∞-norm constraints to craft input perturbations, only a few investigate sparse ℓ_1- and ℓ_0-norm attacks.In particular, ℓ_0-norm attacks remain the least studied due to the inherent complexity of optimizing over a non-convex and non-differentiable constraint.However, evaluating adversarial robustness under these attacks could reveal weaknesses otherwise left untested with more conventional ℓ_2- and ℓ_∞-norm attacks.In this work, we propose a novel ℓ_0-norm attack, called σ-zero, which leverages a differentiable approximation of the ℓ_0 norm to facilitate gradient-based optimization, and an adaptive projection operator to dynamically adjust the trade-off between loss minimization and perturbation sparsity.Extensive evaluations using MNIST, CIFAR10, and ImageNet datasets, involving robust and non-robust models, show that σ-zero finds minimum ℓ_0-norm adversarial examples without requiring any time-consuming hyperparameter tuning, and that it outperforms all competing sparse attacks in terms of success rate, perturbation size, and efficiency.

2016. pMoE: Prompting Diverse Experts Together Wins More in Visual Adaptation

链接：https://iclr.cc/virtual/2025/poster/28112 abstract： Parameter-efficient fine-tuning has demonstrated promising results across various visual adaptation tasks, such as classification and segmentation. Typically, prompt tuning techniques have harnessed knowledge from a single pre-trained model, whether from a general or a specialized medical domain. However, this approach typically overlooks the potential synergies that could arise from integrating diverse domain knowledge within the same tuning process. In this work, we propose a novel Mixture-of-Experts prompt tuning method called pMoE, which leverages the strengths of multiple expert domains through expert-specialized prompt tokens and the learnable dispatcher, effectively combining their expertise in a unified model framework. Our pMoE introduces expert-specific prompt tokens and utilizes a dynamic token dispatching mechanism at various prompt layers to optimize the contribution of each domain expert during the adaptation phase. By incorporating both domain knowledge from diverse experts, the proposed pMoE significantly enhances the model's versatility and applicability to a broad spectrum of tasks. We conduct extensive experiments across 47 adaptation tasks, including both classification and segmentation in general and medical domains. The results demonstrate that our pMoE not only achieves superior performance with a large margin of improvements but also offers an optimal trade-off between computational efficiency and adaptation effectiveness compared to existing methods.

2017. Comparing noisy neural population dynamics using optimal transport distances

链接：https://iclr.cc/virtual/2025/poster/29054 abstract： Biological and artificial neural systems form high-dimensional neural representations that underpin their computational capabilities. Methods for quantifying geometric similarity in neural representations have become a popular tool for identifying computational principles that are potentially shared across neural systems. These methods generally assume that neural responses are deterministic and static. However, responses of biological systems, and some artificial systems, are noisy and dynamically unfold over time. Furthermore, these characteristics can have substantial influence on a system’s computational capabilities. Here, we demonstrate that existing metrics can fail to capture key differences between neural systems with noisy dynamic responses. We then propose a metric for comparing the geometry of noisy neural trajectories, which can be derived as an optimal transport distance between Gaussian processes. We use the metric to compare models of neural responses in different regions of the motor system and to compare the dynamics of latent diffusion models for text-to-image synthesis.

2018. Physics-Informed Deep Inverse Operator Networks for Solving PDE Inverse Problems

链接：https://iclr.cc/virtual/2025/poster/31262 abstract： Inverse problems involving partial differential equations (PDEs) can be seen as discovering a mapping from measurement data to unknown quantities, often framed within an operator learning approach. However, existing methods typically rely on large amounts of labeled training data, which is impractical for most real-world applications. Moreover, these supervised models may fail to capture the underlying physical principles accurately. To address these limitations, we propose a novel architecture called Physics-Informed Deep Inverse Operator Networks (PI-DIONs), which can learn the solution operator of PDE-based inverse problems without any labeled training data. We extend the stability estimates established in the inverse problem literature to the operator learning framework, thereby providing a robust theoretical foundation for our method. These estimates guarantee that the proposed model, trained on a finite sample and grid, generalizes effectively across the entire domain and function space. Extensive experiments are conducted to demonstrate that PI-DIONs can effectively and accurately learn the solution operators of the inverse problems without the need for labeled data.

2019. ALLaM: Large Language Models for Arabic and English

链接：https://iclr.cc/virtual/2025/poster/29915 abstract： In this work, we present ALLaM: Arabic Large Language Model, a series of large language models to support the ecosystem of Arabic Language Technologies (ALT). ALLaM is carefully trained, considering the values of language alignment and transferability of knowledge at scale. The models are based on an autoregressive decoder-only architecture and are pretrained on a mixture of Arabic and English texts. We illustrate how the second-language acquisition via vocabulary expansion can help steer a language model towards a new language without any major catastrophic forgetting in English. Furthermore, we highlight the effectiveness of using translation data and the process of knowledge encoding within the language model's latent space. Finally, we show that effective alignment with human preferences can significantly enhance the performance of a large language model (LLM) compared to less aligned models of a larger scale. Our methodology enables us to achieve state-of-the-art performance in various Arabic benchmarks, including MMLU Arabic, ACVA, and Arabic Exams. Our aligned models improve both in Arabic and English from its base aligned models.

2020. ParFam -- (Neural Guided) Symbolic Regression via Continuous Global Optimization

链接：https://iclr.cc/virtual/2025/poster/30731 abstract： The problem of symbolic regression (SR) arises in many different applications, such as identifying physical laws or deriving mathematical equations describing the behavior of financial markets from given data. Various methods exist to address the problem of SR, often based on genetic programming. However, these methods are usually complicated and involve various hyperparameters. In this paper, we present our new approach ParFam that utilizes parametric families of suitable symbolic functions to translate the discrete symbolic regression problem into a continuous one, resulting in a more straightforward setup compared to current state-of-the-art methods. In combination with a global optimizer, this approach results in a highly effective method to tackle the problem of SR. We theoretically analyze the expressivity of ParFam and demonstrate its performance with extensive numerical experiments based on the common SR benchmark suit SRBench, showing that we achieve state-of-the-art results. Moreover, we present an extension incorporating a pre-trained transformer network (DL-ParFam) to guide ParFam, accelerating the optimization process by up to two magnitudes. Our code and results can be found at https://anonymous.4open.science/r/parfam-D402.

2021. OSDA Agent: Leveraging Large Language Models for De Novo Design of Organic Structure Directing Agents

链接：https://iclr.cc/virtual/2025/poster/30683 abstract： Zeolites are crystalline porous materials that have been widely utilized in petrochemical industries as well as sustainable chemistry areas. Synthesis of zeolites often requires small molecules termed Organic Structure Directing Agents (OSDAs), which are critical in forming the porous structure. Molecule generation models can aid the design of OSDAs, but they are limited by single functionality and lack of interactivity. Meanwhile, large language models (LLMs) such as GPT-4, as general-purpose artificial intelligence systems, excel in instruction comprehension, logical reasoning, and interactive communication. However, LLMs lack in-depth chemistry knowledge and first-principle computation capabilities, resulting in uncontrollable outcomes even after fine-tuning. In this paper, we propose OSDA Agent, an interactive OSDA design framework that leverages LLMs as the brain, coupled with computational chemistry tools. The OSDA Agent consists of three main components: the Actor, responsible for generating potential OSDA structures; the Evaluator, which assesses and scores the generated OSDAs using computational chemistry tools; and the Self-reflector, which produces reflective summaries based on the Evaluator's feedback to refine the Actor's subsequent outputs. Experiments on representative zeolite frameworks show the generation-evaluation-reflection-refinement workflow can perform de novo design of OSDAs with superior generation quality than the pure LLM model, generating candidates consistent with experimentally validated OSDAs and optimizing known OSDAs.

2022. Progressive Compositionality in Text-to-Image Generative Models

链接：https://iclr.cc/virtual/2025/poster/29626 abstract： Despite the impressive text-to-image (T2I) synthesis capabilities of diffusion models, they often struggle to understand compositional relationships between objects and attributes, especially in complex settings. Existing approaches through building compositional architectures or generating difficult negative captions often assume a fixed prespecified compositional structure, which limits generalization to new distributions. In this paper, we argue that curriculum training is crucial to equipping generative models with a fundamental understanding of compositionality. To achieve this, we leverage large-language models (LLMs) to automatically compose complex scenarios and harness Visual-Question Answering (VQA) checkers to automatically curate a contrastive dataset, ConPair, consisting of 15k pairs of high-quality contrastive images. These pairs feature minimal visual discrepancies and cover a wide range of attribute categories, especially complex and natural scenarios. To learn effectively from these error cases (i.e., hard negative images), we propose EvoGen, a new multi-stage curriculum for contrastive learning of diffusion models. Through extensive experiments across a wide range of compositional scenarios, we showcase the effectiveness of our proposed framework on compositional T2I benchmarks.

2023. Analyzing and Boosting the Power of Fine-Grained Visual Recognition for Multi-modal Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28323 abstract： Multi-modal large language models (MLLMs) have shown remarkable abilities in various visual understanding tasks. However, MLLMs still struggle with fine-grained visual recognition (FGVR), which aims to identify subordinate-level categories from images. This can negatively impact more advanced capabilities of MLLMs, such as object-centric visual question answering and reasoning. In our study, we revisit three quintessential capabilities of MLLMs for FGVR, including object information extraction, category knowledge reserve, object-category alignment, and position of the root cause as a misalignment problem. To address this issue, we present Finedefics, an MLLM that enhances the model's FGVR capability by incorporating informative attribute descriptions of objects into the training phase. We employ contrastive learning on object-attribute pairs and attribute-category pairs simultaneously and use examples from similar but incorrect categories as hard negatives, naturally bringing representations of visual objects and category names closer. Extensive evaluations across multiple popular FGVR datasets demonstrate that Finedefics outperforms existing MLLMs of comparable parameter sizes, showcasing its remarkable efficacy. The code is available at https://github.com/PKU-ICST-MIPL/Finedefics_ICLR2025.

2024. Learning to engineer protein flexibility

链接：https://iclr.cc/virtual/2025/poster/30019 abstract： Generative machine learning models are increasingly being used to design novel proteins. However, their major limitation is the inability to account for protein flexibility, a property crucial for protein function. Learning to engineer flexibility is difficult because the relevant data is scarce, heterogeneous, and costly to obtain using computational and experimental methods. Our contributions are three-fold. First, we perform a comprehensive comparison of methods for evaluating protein flexibility and identify relevant data for learning. Second, we overcome the data scarcity issue by leveraging a pre-trained protein language model. We design and train flexibility predictors utilizing either only sequential or both sequential and structural information on the input. Third, we introduce a method for fine-tuning a protein inverse folding model to make it steerable toward desired flexibility at specified regions. We demonstrate that our method Flexpert enables guidance of inverse folding models toward increased flexibility. This opens up a transformative possibility of engineering protein flexibility.

2025. OmniPhysGS: 3D Constitutive Gaussians for General Physics-Based Dynamics Generation

链接：https://iclr.cc/virtual/2025/poster/30706 abstract： Recently, significant advancements have been made in the reconstruction and generation of 3D assets, including static cases and those with physical interactions. To recover the physical properties of 3D assets, existing methods typically assume that all materials belong to a specific predefined category (e.g., elasticity). However, such assumptions ignore the complex composition of multiple heterogeneous objects in real scenarios and tend to render less physically plausible animation given a wider range of objects. We propose OmniPhysGS for synthesizing a physics-based 3D dynamic scene composed of more general objects. A key design of OmniPhysGS is treating each 3D asset as a collection of constitutive 3D Gaussians. For each Gaussian, its physical material is represented by an ensemble of 12 physical domain-expert sub-models (rubber, metal, honey, water, etc.), which greatly enhances the flexibility of the proposed model. In the implementation, we define a scene by user-specified prompts and supervise the estimation of material weighting factors via a pretrained video diffusion model. Comprehensive experiments demonstrate that OmniPhysGS achieves more general and realistic physical dynamics across a broader spectrum of materials, including elastic, viscoelastic, plastic, and fluid substances, as well as interactions between different materials. Our method surpasses existing methods by approximately 3% to 16% in metrics of visual quality and text alignment.

2026. Polyrating: A Cost-Effective and Bias-Aware Rating System for LLM Evaluation

链接：https://iclr.cc/virtual/2025/poster/29474 abstract： Rating-based human evaluation has become an essential tool to accurately evaluate the impressive performance of large language models (LLMs). However, current rating systems suffer from several important limitations: first, they fail to account for biases that significantly influence evaluation results, second, they require large and expensive preference datasets to obtain accurate ratings, and third, they do not facilitate meaningful comparisons of model ratings across different tasks. To address these issues, we introduce Polyrating, an expressive and flexible rating system based on maximum a posteriori estimation that enables a more nuanced and thorough analysis of model performance at lower costs. Polyrating can detect and quantify biases affecting human preferences, ensuring fairer model comparisons. Further, Polyrating can reduce the cost of human evaluations by up to 41% for new models and up to 77% for new tasks by leveraging existing benchmark scores. Lastly, Polyrating enables direct comparisons of ratings across different tasks, providing a comprehensive understanding of an LLMs' strengths, weaknesses, and relative performance across different applications.

2027. ACES: Automatic Cohort Extraction System for Event-Stream Datasets

链接：https://iclr.cc/virtual/2025/poster/29780 abstract： Reproducibility remains a significant challenge in machine learning (ML) for healthcare. Datasets, model pipelines, and even task or cohort definitions are often private in this field, leading to a significant barrier in sharing, iterating, and understanding ML results on electronic health record (EHR) datasets. We address a significant part of this problem by introducing the Automatic Cohort Extraction System (ACES) for event-stream data. This library is designed to simultaneously simplify the development of tasks and cohorts for ML in healthcare and also enable their reproduction, both at an exact level for single datasets and at a conceptual level across datasets. To accomplish this, ACES provides: (1) a highly intuitive and expressive domain-specific configuration language for defining both dataset-specific concepts and dataset-agnostic inclusion or exclusion criteria, and (2) a pipeline to automatically extract patient records that meet these defined criteria from real-world data. ACES can be automatically applied to any dataset in either the Medical Event Data Standard (MEDS) or Event Stream GPT (ESGPT) formats, or to any dataset in which the necessary task-specific predicates can be extracted in an event-stream form. ACES has the potential to significantly lower the barrier to entry for defining ML tasks in representation learning, redefine the way researchers interact with EHR datasets, and significantly improve the state of reproducibility for ML studies using this modality. ACES is available at: https://github.com/justin13601/aces.

2028. Learning multi-modal generative models with permutation-invariant encoders and tighter variational objectives

链接：https://iclr.cc/virtual/2025/poster/31466 abstract： Devising deep latent variable models for multi-modal data has been a long-standing theme in machine learning research. Multi-modal Variational Autoencoders (VAEs) have been a popular generative model class that learns latent representations that jointly explain multiple modalities. Various objective functions for such models have been suggested, often motivated as lower bounds on the multi-modal data log-likelihood or from information-theoretic considerations. To encode latent variables from different modality subsets, Product-of-Experts (PoE) or Mixture-of-Experts (MoE) aggregation schemes have been routinely used and shown to yield different trade-offs, for instance, regarding their generative quality or consistency across multiple modalities. In this work, we consider a variational objective that can tightly approximate the data log-likelihood. We develop more flexible aggregation schemes that avoid the inductive biases in PoE or MoE approaches by combining encoded features from different modalities based on permutation-invariant neural networks. Our numerical experiments illustrate trade-offs for multi-modal variational objectives and various aggregation schemes. We show that our variational objective and more flexible aggregation models can become beneficial when one wants to approximate the true joint distribution over observed modalities and latent variables in identifiable models.

2029. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning

链接：https://iclr.cc/virtual/2025/poster/27873 abstract： Chain-of-thought (CoT) via prompting is the de facto method for eliciting reasoning capabilities from large language models (LLMs). But for what kinds of tasks is this extra "thinking" really helpful? To analyze this, we conducted a quantitative meta-analysis covering over 100 papers using CoT and ran our own evaluations of 20 datasets across 14 models. Our results show that CoT gives strong performance benefits primarily on tasks involving math or logic, with much smaller gains on other types of tasks. On MMLU, directly generating the answer without CoT leads to almost identical accuracy as CoT unless the question or model's response contains an equals sign, indicating symbolic operations and reasoning. Following this finding, we analyze the behavior of CoT on these problems by separating planning and execution and comparing against tool-augmented LLMs. Much of CoT's gain comes from improving symbolic execution, but it underperforms relative to using a symbolic solver. Our results indicate that CoT can be applied selectively, maintaining performance while saving inference costs. Furthermore, they suggest a need to move beyond prompt-based CoT to new paradigms that better leverage intermediate computation across the whole range of LLM applications.

2030. RDT-1B: a Diffusion Foundation Model for Bimanual Manipulation

链接：https://iclr.cc/virtual/2025/poster/27746 abstract： Bimanual manipulation is essential in robotics, yet developing foundation models is extremely challenging due to the inherent complexity of coordinating two robot arms (leading to multi-modal action distributions) and the scarcity of training data. In this paper, we present the Robotics Diffusion Transformer (RDT), a pioneering diffusion foundation model for bimanual manipulation. RDT builds on diffusion models to effectively represent multi-modality, with innovative designs of a scalable Transformer to deal with the heterogeneity of multi-modal inputs and to capture the nonlinearity and high frequency of robotic data. To address data scarcity, we further introduce a Physically Interpretable Unified Action Space, which can unify the action representations of various robots while preserving the physical meanings of original actions, facilitating learning transferrable physical knowledge. With these designs, we managed to pre-train RDT on the largest collection of multi-robot datasets to date and scaled it up to 1.2B parameters, which is the largest diffusion-based foundation model for robotic manipulation. We finally fine-tuned RDT on a self-created multi-task bimanual dataset with over 6K+ episodes to refine its manipulation capabilities. Experiments on real robots demonstrate that RDT significantly outperforms existing methods. It exhibits zero-shot generalization to unseen objects and scenes, understands and follows language instructions, learns new skills with just 1$\sim$5 demonstrations, and effectively handles complex, dexterous tasks. We refer to https://rdt-robotics.github.io/rdt-robotics/ for the code and videos.

2031. GRAIN: Exact Graph Reconstruction from Gradients

链接：https://iclr.cc/virtual/2025/poster/30812 abstract： Federated learning claims to enable collaborative model training among multiple clients with data privacy by transmitting gradient updates instead of the actual client data. However, recent studies have shown the client privacy is still at risk due to the, so called, gradient inversion attacks which can precisely reconstruct clients' text and image data from the shared gradient updates. While these attacks demonstrate severe privacy risks for certain domains and architectures, the vulnerability of other commonly-used data types, such as graph-structured data, remain under-explored. To bridge this gap, we present GRAIN, the first exact gradient inversion attack on graph data in the honest-but-curious setting that recovers both the structure of the graph and the associated node features. Concretely, we focus on Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT) -- two of the most widely used frameworks for learning on graphs. Our method first utilizes the low-rank structure of GNN gradients to efficiently reconstruct and filter the client subgraphs which are then joined to complete the input graph. We evaluate our approach on molecular, citation, and social network datasets using our novel metric. We show that GRAIN reconstructs up to 80\% of all graphs exactly, significantly outperforming the baseline, which achieves up to 20\% correctly positioned nodes.

2032. Synergy and Diversity in CLIP: Enhancing Performance Through Adaptive Backbone Ensembling

链接：https://iclr.cc/virtual/2025/poster/29198 abstract： Contrastive Language-Image Pretraining (CLIP) stands out as a prominent method for image representation learning. Various architectures, from vision transformers~(ViTs) to convolutional networks (ResNets) have been trained with CLIP to serve as general solutions to diverse vision tasks.This paper explores the differences across various CLIP-trained vision backbones.Despite using the same data and training objective, we find that these architectures have notably different representations,different classification performance across datasets, and different robustness properties to certain types of image perturbations.Our findings indicate a remarkable possible synergy across backbonesby leveraging their respective strengths.In principle, classification accuracy could be improved by over 40 percentage with an informed selection of the optimal backbone per test example. Using this insight, we develop a straightforward yet powerful approach to adaptively ensemble multiple backbones.The approach uses as few as one labeled example per classto tune the adaptive combination of backbones.On a large collection of datasets, the method achieves a remarkable increase in accuracy of up to 39.1\% over the best single backbone, well beyond traditional ensembles.

2033. Diff-2-in-1: Bridging Generation and Dense Perception with Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29212 abstract： Beyond high-fidelity image synthesis, diffusion models have recently exhibited promising results in dense visual perception tasks. However, most existing work treats diffusion models as a standalone component for perception tasks, employing them either solely for off-the-shelf data augmentation or as mere feature extractors. In contrast to these isolated and thus sub-optimal efforts, we introduce an integrated, versatile, diffusion-based framework, Diff-2-in-1, that can simultaneously handle both multi-modal data generation and dense visual perception, through a unique exploitation of the diffusion-denoising process. Within this framework, we further enhance discriminative visual perception via multi-modal generation, by utilizing the denoising network to create multi-modal data that mirror the distribution of the original training set. Importantly, Diff-2-in-1 optimizes the utilization of the created diverse and faithful data by leveraging a novel self-improving learning mechanism. Comprehensive experimental evaluations validate the effectiveness of our framework, showcasing consistent performance improvements across various discriminative backbones and high-quality multi-modal data generation characterized by both realism and usefulness. Our project website is available at https://zsh2000.github.io/diff-2-in-1.github.io/.

2034. Gradient descent with generalized Newton’s method

链接：https://iclr.cc/virtual/2025/poster/29116 abstract： We propose the generalized Newton's method (GeN) --- a Hessian-informed approach that applies to any optimizer such as SGD and Adam, and covers the Newton-Raphson method as a sub-case. Our method automatically and dynamically selects the learning rate that accelerates the convergence, without the intensive tuning of the learning rate scheduler. In practice, our method is easily implementable, since it only requires additional forward passes with almost zero computational overhead (in terms of training time and memory cost), if the overhead is amortized over many iterations. We present extensive experiments on language and vision tasks (e.g. GPT and ResNet) to showcase that GeN optimizers match the state-of-the-art performance, which was achieved with carefully tuned learning rate schedulers.

2035. Methods for Convex (L_0,L_1)-Smooth Optimization: Clipping, Acceleration, and Adaptivity

链接：https://iclr.cc/virtual/2025/poster/31230 abstract： Due to the non-smoothness of optimization problems in Machine Learning, generalized smoothness assumptions have been gaining a lot of attention in recent years. One of the most popular assumptions of this type is (L_0,L_1)-smoothness (Zhang et al., 2020). In this paper, we focus on the class of (strongly) convex (L_0,L_1)-smooth functions and derive new convergence guarantees for several existing methods. In particular, we derive improved convergence rates for Gradient Descent with (Smoothed) Gradient Clipping and for Gradient Descent with Polyak Stepsizes. In contrast to the existing results, our rates do not rely on the standard smoothness assumption and do not suffer from the exponential dependency on the initial distance to the solution. We also extend these results to the stochastic case under the over-parameterization assumption, propose a new accelerated method for convex (L_0,L_1)-smooth optimization, and derive new convergence rates for Adaptive Gradient Descent (Malitsky and Mishchenko, 2020).

2036. Exposing and Addressing Cross-Task Inconsistency in Unified Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/31504 abstract： As general purpose vision models get increasingly effective at a wide set of tasks, it is imperative that they be consistent across the tasks they support. Inconsistent AI models are considered brittle and untrustworthy by human users and are more challenging to incorporate into larger systems that take dependencies on their outputs. Measuring consistency between very heterogeneous tasks that might include outputs in different modalities is challenging since it is difficult to determine if the predictions are consistent with one another. As a solution, we introduce a benchmark dataset, CocoCON, where we create contrast sets by modifying test instances for multiple tasks in small but semantically meaningful ways to change the gold label and outline metrics for measuring if a model is consistent by ranking the original and perturbed instances across tasks. We find that state-of-the-art vision-language models suffer from a surprisingly high degree of inconsistent behavior across tasks, especially for more heterogeneous tasks. To alleviate this issue, we propose a rank correlation-based auxiliary training objective, computed over large automatically created cross-task contrast sets, that improves the multi-task consistency of large unified models while retaining their original accuracy on downstream tasks.

2037. PeriodWave: Multi-Period Flow Matching for High-Fidelity Waveform Generation

链接：https://iclr.cc/virtual/2025/poster/28055 abstract： Recently, universal waveform generation tasks have been investigated conditioned on various out-of-distribution scenarios. Although one-step GAN-based methods have shown their strength in fast waveform generation, they are vulnerable to train-inference mismatch scenarios such as two-stage text-to-speech. Meanwhile, diffusion-based models have shown their powerful generative performance in other domains; however, they stay out of the limelight due to slow inference speed in waveform generation tasks. Above all, there is no generator architecture that can explicitly disentangle the natural periodic features of high-resolution waveform signals. In this paper, we propose PeriodWave, a novel universal waveform generation model from Mel-spectrogram and neural audio codec. First, we introduce a period-aware flow matching estimator that effectively captures the periodic features of the waveform signal when estimating the vector fields. Additionally, we utilize a multi-period estimator that avoids overlaps to capture different periodic features of waveform signals. Although increasing the number of periods can improve the performance significantly, this requires more computational costs. To reduce this issue, we also propose a single period-conditional universal estimator that can feed-forward parallel by period-wise batch inference. Additionally, we first introduce FreeU to reduce the high-frequency noise for waveform generation. Furthermore, we demonstrate the effectiveness of the proposed method in neural audio codec decoding task, and present the streaming generation framework of non-autoregressive model for speech language models. The experimental results demonstrated that our model outperforms the previous models in reconstruction tasks from Mel-spectrogram and discrete token, and text-to-speech tasks. Source code is available at https://github.com/sh-lee-prml/PeriodWave

2038. Nesterov acceleration in benignly non-convex landscapes

链接：https://iclr.cc/virtual/2025/poster/29241 abstract： While momentum-based optimization algorithms are commonly used in the notoriously non-convex optimization problems of deep learning, their analysis has historically been restricted to the convex and strongly convex setting. In this article, we partially close this gap between theory and practice and demonstrate that virtually identical guarantees can be obtained in optimization problems with a 'benign' non-convexity. We show that these weaker geometric assumptions are well justified in overparametrized deep learning, at least locally. Variations of this result are obtained for a continuous time model of Nesterov's accelerated gradient descent algorithm (NAG), the classical discrete time version of NAG, and versions of NAG with stochastic gradient estimates with purely additive noise and with noise that exhibits both additive and multiplicative scaling.

2039. Towards hyperparameter-free optimization with differential privacy

链接：https://iclr.cc/virtual/2025/poster/31119 abstract： Differential privacy (DP) is a privacy-preserving paradigm that protects the training data when training deep learning models. Critically, the performance of models is determined by the training hyperparameters, especially those of the learning rate schedule, thus requiring fine-grained hyperparameter tuning on the data. In practice, it is common to tune the learning rate hyperparameters through the grid search that (1) is computationally expensive as multiple runs are needed, and (2) increases the risk of data leakage as the selection of hyperparameters is data-dependent. In this work, we adapt the automatic learning rate schedule to DP optimization for any models and optimizers, so as to significantly mitigate or even eliminate the cost of hyperparameter tuning when applied together with automatic per-sample gradient clipping. Our hyperparameter-free DP optimization is almost as computationally efficient as the standard non-DP optimization, and achieves state-of-the-art DP performance on various language and vision tasks.

2040. MMAD: A Comprehensive Benchmark for Multimodal Large Language Models in Industrial Anomaly Detection

链接：https://iclr.cc/virtual/2025/poster/30120 abstract： In the field of industrial inspection, Multimodal Large Language Models (MLLMs) have a high potential to renew the paradigms in practical applications due to their robust language capabilities and generalization abilities. However, despite their impressive problem-solving skills in many domains, MLLMs' ability in industrial anomaly detection has not been systematically studied. To bridge this gap, we present MMAD, a full-spectrum MLLM benchmark in industrial Anomaly Detection. We defined seven key subtasks of MLLMs in industrial inspection and designed a novel pipeline to generate the MMAD dataset with 39,672 questions for 8,366 industrial images. With MMAD, we have conducted a comprehensive, quantitative evaluation of various state-of-the-art MLLMs. The commercial models performed the best, with the average accuracy of GPT-4o models reaching 74.9\%. However, this result falls far short of industrial requirements. Our analysis reveals that current MLLMs still have significant room for improvement in answering questions related to industrial anomalies and defects. We further explore two training-free performance enhancement strategies to help models improve in industrial scenarios, highlighting their promising potential for future research. The code and data are available at https://github.com/jam-cc/MMAD.

2041. See It from My Perspective: How Language Affects Cultural Bias in Image Understanding

链接：https://iclr.cc/virtual/2025/poster/29299 abstract： Vision-language models (VLMs) can respond to queries about images in many languages. However, beyond language, culture affects how we see things. For example, individuals from Western cultures focus more on the central figure in an image while individuals from East Asian cultures attend more to scene context (Nisbett 2001). In this work, we characterize the Western bias of VLMs in image understanding and investigate the role that language plays in this disparity. We evaluate VLMs across subjective and objective visual tasks with culturally diverse images and annotations. We find that VLMs perform better on the Western split than on the East Asian split of each task. Through controlled experimentation, we trace one source of this bias in image understanding to the lack of diversity in language model construction. While inference in a language nearer to a culture can lead to reductions in bias, we show it is much more effective when that language was well-represented during text-only pre-training. Interestingly, this yields bias reductions even when prompting in English. Our work highlights the importance of richer representation of all languages in building equitable VLMs.

2042. LevAttention: Time, Space and Streaming Efficient Algorithm for Heavy Attentions

链接：https://iclr.cc/virtual/2025/poster/30728 abstract： A central problem related to transformers can be stated as follows: given two $n \times d$ matrices Q and K, and a non-negative function f, define the matrix A as follows: (1) apply the function f to each entry of the $n \times n$ matrix $Q K^T$, and then (2) normalize each of the row sums of A to be equal to 1. The matrix A can be computed in $O(n^2 d)$ time assuming f can be applied to a number in constant time, but the quadratic dependence on n is prohibitive in applications where it corresponds to long context lengths. For a large class of functions f, we show how to find all the "large attention scores", i.e., entries of A which are at least a positive value ε, in time with linear dependence on n (i.e., $n \cdot \textrm{poly}(d/\varepsilon)$) for a positive parameter $\varepsilon > 0$. Our class of functions include all functions f of the form $f(x) = |x|^p$, as explored recently in transformer models. Using recently developed tools from randomized numerical linear algebra, we prove that for any K, there is a "universal set" $U \subset [n]$ of size independent of n, such that for any Q and any row i, the large attention scores $A_{i,j}$ in row i of A all have $j \in U$. We also find U in $n \cdot \textrm{poly}(d/\varepsilon)$ time. Notably, we (1) make no assumptions on the data, (2) our workspace does not grow with n, and (3) our algorithms can be computed in streaming and parallel settings. We empirically show the benefits of our scheme for vision transformers, showing how to train new models that use our universal set while training as well, showing that our model is able to consistently select "important keys'" during training. We also provide theoretical motivation by formulating a planted model in which our efficient algorithms provably identify relevant keys for each query.

2043. Ctrl-Adapter: An Efficient and Versatile Framework for Adapting Diverse Controls to Any Diffusion Model

链接：https://iclr.cc/virtual/2025/poster/28385 abstract： ControlNets are widely used for adding spatial control to text-to-image diffusion models. However, when it comes to controllable video generation, ControlNets cannot be directly integrated into new backbones due to feature space mismatches, and training ControlNets for new backbones can be a significant burden for many users. Furthermore, applying ControlNets independently to different frames can not effectively maintain object temporal consistency. To address these challenges, we introduce Ctrl-Adapter, an efficient and versatile framework that adds diverse controls to any image/video diffusion models through the adaptation of pretrained ControlNets. Ctrl-Adapter offers strong and diverse capabilities, including image and video control, sparse-frame video control, fine-grained patch-level multi-condition control, zero-shot adaptation to unseen conditions, and supports a variety of downstream tasks beyond spatial control, including video editing, video style transfer, and text-guided motion control. With six diverse U-Net/DiT-based image/video diffusion models (SDXL, PixArt-α, I2VGen-XL, SVD, Latte, Hotshot-XL), Ctrl-Adapter matches the performance of pretrained ControlNets on COCO and achieves the state-of-the-art on DAVIS 2017 with significantly lower computation (< 10 GPU hours).

2044. From Attention to Activation: Unraveling the Enigmas of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30153 abstract： We study two strange phenomena in auto-regressive Transformers: (1) the dominance of the first token in attention heads; (2) the occurrence of large outlier activations in the hidden states. We find that popular large language models, such as Llama attend maximally to the first token in 98% of attention heads, a behaviour we attribute to the softmax function. To mitigate this issue, we propose a reformulation of softmax to softmax-1. Furthermore, we identify adaptive optimisers, e.g. Adam, as the primary contributor to the large outlier activations and introduce OrthoAdam, a novel optimiser that utilises orthogonal matrices to transform gradients, to address this issue. Finally, not only do our methods prevent these phenomena from occurring, but additionally, they enable Transformers to sustain their performance when quantised using basic algorithms, something that standard methods are unable to do. In summary, our methods reduce the attention proportion on the first token from 65% to 3.3%, the activation kurtosis in the hidden states from 1657 to 3.1, and perplexity penalty under 4-bit weight quantisation from 3565 to 0.3. Code is available at https://github.com/prannaykaul/OrthoAdam

2045. Generative Inbetweening: Adapting Image-to-Video Models for Keyframe Interpolation

链接：https://iclr.cc/virtual/2025/poster/27704 abstract： We present a method for generating video sequences with coherent motion between a pair of input keyframes. We adapt a pretrained large-scale image-to-video diffusion model (originally trained to generate videos moving forward in time from a single input image) for keyframe interpolation, i.e., to produce a video between two input frames. We accomplish this adaptation through a lightweight fine-tuning technique that produces a version of the model that instead predicts videos moving backwards in time from a single input image. This model (along with the original forward-moving model) is subsequently used in a dual-directional diffusion sampling process that combines the overlapping model estimates starting from each of the two keyframes. Our experiments shows that our method outperforms both existing diffusion-based methods and traditional frame interpolation techniques.

2046. Long-time asymptotics of noisy SVGD outside the population limit

链接：https://iclr.cc/virtual/2025/poster/29322 abstract： Stein Variational Gradient Descent (SVGD) is a widely used sampling algorithm that has been successfully applied in several areas of Machine Learning. SVGD operates by iteratively moving a set of n interacting particles (which represent the samples) to approximate the target distribution. Despite recent studies on the complexity of SVGD and its variants, their long-time asymptotic behavior (i.e., after numerous iterations k) is still not understood in the finite number of particles regime. We study the long-time asymptotic behavior of a noisy variant of SVGD. First, we establish that the limit set of noisy SVGD for large k is well-defined. We then characterize this limit set, showing that it approaches the target distribution as n increases. In particular, noisy SVGD avoids the variance collapse observed for

SVGD. Our approach involves demonstrating that the trajectories of noisy SVGD closely resemble those described by a McKean-Vlasov process.

2047. MamKO: Mamba-based Koopman operator for modeling and predictive control

链接：https://iclr.cc/virtual/2025/poster/28767 abstract： The Koopman theory, which enables the transformation of nonlinear systems into linear representations, is a powerful and efficient tool to model and control nonlinear systems. However, the ability of the Koopman operator to model complex systems, particularly time-varying systems, is limited by the fixed linear state-space representation. To address the limitation, the large language model, Mamba, is considered a promising strategy for enhancing modeling capabilities while preserving the linear state-space structure.In this paper, we propose a new framework, the Mamba-based Koopman operator (MamKO), which provides enhanced model prediction capability and adaptability, as compared to Koopman models with constant Koopman operators. Inspired by the Mamba structure, MamKO generates Koopman operators from online data; this enables the model to effectively capture the dynamic behaviors of the nonlinear system over time. A model predictive control system is then developed based on the proposed MamKO model. The modeling and control performance of the proposed method is evaluated through experiments on benchmark time-invariant and time-varying systems. The experimental results demonstrate the superiority of the proposed approach. Additionally, we perform ablation experiments to test the effectiveness of individual components of MamKO. This approach unlocks new possibilities for integrating large language models with control frameworks, and it achieves a good balance between advanced modeling capabilities and real-time control implementation efficiency.

2048. FlexPrefill: A Context-Aware Sparse Attention Mechanism for Efficient Long-Sequence Inference

链接：https://iclr.cc/virtual/2025/poster/29811 abstract： Large language models (LLMs) encounter computational challenges during long-sequence inference, especially in the attention pre-filling phase, where the complexity grows quadratically with the prompt length. Previous efforts to mitigate these challenges have relied on fixed sparse attention patterns or identifying sparse attention patterns based on limited cases. However, these methods lacked the flexibility to efficiently adapt to varying input demands. In this paper, we introduce FlexPrefill, a Flexible sparse Pre-filling mechanism that dynamically adjusts sparse attention patterns and computational budget in real-time to meet the specific requirements of each input and attention head. The flexibility of our method is demonstrated through two key innovations: 1) Query-Aware Sparse Pattern Determination: By measuring Jensen-Shannon divergence, this component adaptively switches between query-specific diverse attention patterns and predefined attention patterns. 2) Cumulative-Attention Based Index Selection: This component dynamically selects query-key indexes to be computed based on different attention patterns, ensuring the sum of attention scores meets a predefined threshold.FlexPrefill adaptively optimizes the sparse pattern and sparse ratio of each attention head based on the prompt, enhancing efficiency in long-sequence inference tasks. Experimental results show significant improvements in both speed and accuracy over prior methods, providing a more flexible and efficient solution for LLM inference.

2049. Streaming Algorithms For ℓ_p Flows and ℓ_p Regression

链接：https://iclr.cc/virtual/2025/poster/30029 abstract： We initiate the study of one-pass streaming algorithms for underdetermined ℓ_p linear regression problems of the form $$ \min_{\mathbf A\mathbf x = \mathbf b} \lVert\mathbf x\rVert_p \,, \qquad \text{where } \mathbf A \in \mathbb R^{n \times d} \text{ with } n \ll d \,, $$ which generalizes basis pursuit ($p = 1$) and least squares solutions to underdetermined linear systems ($p = 2$). We study the column-arrival streaming model, in which the columns of $\mathbf A$ are presented one by one in a stream. When $\mathbf A$ is the incidence matrix of a graph, this corresponds to an edge insertion graph stream, and the regression problem captures ℓ_p flows which includes transshipment ($p = 1$), electrical flows ($p = 2$), and max flow ($p = \infty$) on undirected graphs as special cases. Our goal is to design algorithms which use space much less than the entire stream, which has a length of d. For the task of estimating the cost of the ℓ_p regression problem for $p\in[2,\infty]$, we show a streaming algorithm which constructs a sparse instance supported on $\tilde O(\varepsilon^{-2}n)$ columns of $\mathbf A$ which approximates the cost up to a $(1\pm\varepsilon)$ factor, which corresponds to $\tilde O(\varepsilon^{-2}n^2)$ bits of space in general and an $\tilde O(\varepsilon^{-2}n)$ space semi-streaming algorithm for constructing ℓ_p flow sparsifiers on graphs. This extends to $p\in(1, 2)$ with $\tilde O(\varepsilon^{2}n^{q/2})$ columns, where q is the H\"older conjugate exponent of p. For $p = 2$, we show that $\Omega(n^2)$ bits of space are required in general even for outputting a constant factor solution. For $p = 1$, we show that the cost cannot be estimated even to an $o(\sqrt n)$ factor in $\mathrm{poly}(n)$ space. On the other hand, if we are interested in outputting a solution $\mathbf x$, then we show that $(1+\varepsilon)$-approximations require $\Omega(d)$ space for $p > 1$, and in general, κ-approximations require $\tilde\Omega(d/\kappa^{2q})$ space for $p > 1$. We complement these lower bounds with the first sublinear space upper bounds for this problem, showing that we can output a κ-approximation using space only $\mathrm{poly}(n) \cdot \tilde O(d/\kappa^q)$ for $p > 1$, as well as a $\sqrt n$-approximation using $\mathrm{poly}(n, \log d)$ space for $p = 1$.

2050. MLLMs Know Where to Look: Training-free Perception of Small Visual Details with Multimodal LLMs

链接：https://iclr.cc/virtual/2025/poster/30449 abstract： Multimodal Large Language Models (MLLMs) have experienced rapid progress in visual recognition tasks in recent years. Given their potential integration into many critical applications, it is important to understand the limitations of their visual perception. In this work, we study whether MLLMs can perceive small visual details as effectively as large ones when answering questions about images. We observe that their performance is very sensitive to the size of the visual subject of the question, and further show that this effect is in fact causal by conducting an intervention study. Next, we study the attention patterns of MLLMs when answering visual questions, and intriguingly find that they consistently know where to look, even when they provide the wrong answer. Based on these findings, we then propose training-free visual intervention methods that leverage the internal knowledge of any MLLM itself, in the form of attention and gradient maps, to enhance its perception of small visual details. We evaluate our proposed methods on two widely-used MLLMs and seven visual question answering benchmarks and show that they can significantly improve MLLMs' accuracy without requiring any training. Our results elucidate the risk of applying MLLMs to visual recognition tasks concerning small details and indicate that visual intervention using the model's internal state is a promising direction to mitigate this risk. Our code is available at: https://github.com/saccharomycetes/mllms_know.

2051. Improving the Sparse Structure Learning of Spiking Neural Networks from the View of Compression Efficiency

链接：https://iclr.cc/virtual/2025/poster/28809 abstract： The human brain utilizes spikes for information transmission and dynamically reorganizes its network structure to boost energy efficiency and cognitive capabilities throughout its lifespan. Drawing inspiration from this spike-based computation, Spiking Neural Networks (SNNs) have been developed to construct event-driven models that emulate this efficiency. Despite these advances, deep SNNs continue to suffer from over-parameterization during training and inference, a stark contrast to the brain’s ability to self-organize. Furthermore, existing sparse SNNs are challenged by maintaining optimal pruning levels due to a static pruning ratio, resulting in either under or over-pruning.In this paper, we propose a novel two-stage dynamic structure learning approach for deep SNNs, aimed at maintaining effective sparse training from scratch while optimizing compression efficiency. The first stage evaluates the compressibility of existing sparse subnetworks within SNNs using the PQ index, which facilitates an adaptive determination of the rewiring ratio for synaptic connections based on data compression insights. In the second stage, this rewiring ratio critically informs the dynamic synaptic connection rewiring process, including both pruning and regrowth. This approach significantly improves the exploration of sparse structures training in deep SNNs, adapting sparsity dynamically from the point view of compression efficiency.Our experiments demonstrate that this sparse training approach not only aligns with the performance of current deep SNNs models but also significantly improves the efficiency of compressing sparse SNNs. Crucially, it preserves the advantages of initiating training with sparse models and offers a promising solution for implementing Edge AI on neuromorphic hardware.

2052. A Theory for Token-Level Harmonization in Retrieval-Augmented Generation

链接：https://iclr.cc/virtual/2025/poster/28042 abstract： Retrieval-augmented generation (RAG) utilizes retrieved texts to enhance large language models (LLMs). Studies show that while RAG provides valuable external information (benefit), it may also mislead LLMs (detriment) with noisy or incorrect retrieved texts. Although many existing methods attempt to preserve benefit and avoid detriment, they lack a theoretical explanation for RAG. The benefit and detriment in the next token prediction of RAG remain a 'black box' that cannot be quantified or compared in an explainable manner, so existing methods are data-driven, need additional utility evaluators or post-hoc. This paper takes the first step towards providing a theory to explain and trade off the benefit and detriment in RAG. We model RAG as the fusion between distributions of LLMs’ knowledge and distributions of retrieved texts. Then, we formalize the trade-off between the value of external knowledge (benefit) and its potential risk of misleading LLMs (detriment) in next token prediction of RAG by distribution difference in this fusion. Finally, we prove that the actual effect of RAG on the token, which is the comparison between benefit and detriment, can be predicted without any training or accessing the utility of retrieval. Based on our theory, we propose a practical novel method, Tok-RAG, which achieves collaborative generation between the pure LLM and RAG at token level to preserve benefit and avoid detriment. Experiments in real-world tasks using LLMs such as OPT, LLaMA-2, and Mistral show the effectiveness of our method and support our theoretical findings. Code is in supplemental material and will be released on GitHub after acceptance.

2053. Emergence of a High-Dimensional Abstraction Phase in Language Transformers

链接：https://iclr.cc/virtual/2025/poster/31244 abstract： A language model (LM) is a mapping from a linguistic context to an output token. However, much remains to be known about this mapping, including how its geometric properties relate to its function. We take a high-level geometric approach to its analysis, observing, across five pre-trained transformer-based LMs and three input datasets, a distinct phase characterized by high intrinsic dimensionality. During this phase, representations (1) correspond to the first full linguistic abstraction of the input; (2) are the first to viably transfer to downstream tasks; (3) predict each other across different LMs. Moreover, we find that an earlier onset of the phase strongly predicts better language modelling performance. In short, our results suggest that a central high-dimensionality phase underlies core linguistic processing in many common LM architectures.

2054. Model merging with SVD to tie the Knots

链接：https://iclr.cc/virtual/2025/poster/30908 abstract： Recent model merging methods demonstrate that the parameters of fully-finetuned models specializing in distinct tasks can be combined into one model capable of solving all tasks without retraining. Yet, this success does not transfer well when merging LoRA finetuned models. We study this phenomenon and observe that the weights of LoRA finetuned models showcase a lower degree of alignment compared to their fully-finetuned counterparts. We hypothesize that improving this alignment is key to obtaining better LoRA model merges, and propose KnOTS to address this problem. KnOTS uses the SVD to jointly transform the weights of different LoRA models into an aligned space, where existing merging methods can be applied. In addition, we introduce a new benchmark that explicitly evaluates whether merged models are general models. Notably, KnOTS consistently improves LoRA merging by up to 4.3% across several vision and language benchmarks, including our new setting. We release our code at: https://github.com/gstoica27/KnOTS.

2055. Temporal Reasoning Transfer from Text to Video

链接：https://iclr.cc/virtual/2025/poster/28136 abstract： Video Large Language Models (Video LLMs) have shown promising capabilities in video comprehension, yet they struggle with tracking temporal changes and reasoning about temporal relationships.While previous research attributed this limitation to the ineffective temporal encoding of visual inputs, our diagnostic study reveals that video representations contain sufficient information for even small probing classifiers to achieve perfect accuracy.Surprisingly, we find that the key bottleneck in Video LLMs' temporal reasoning capability stems from the underlying LLM's inherent difficulty with temporal concepts, as evidenced by poor performance on textual temporal question-answering tasks.Building on this discovery, we introduce the Textual Temporal reasoning Transfer (T3). T3 synthesizes diverse temporal reasoning tasks in pure text format from existing image-text datasets, addressing the scarcity of video samples with complex temporal scenarios. Remarkably, without using any video data, T3 enhances LongVA-7B's temporal understanding, yielding a 5.3 absolute accuracy improvement on the challenging TempCompass benchmark, which enables our model to outperform ShareGPT4Video-8B trained on 28,000 video samples.Additionally, the enhanced LongVA-7B model achieves competitive performance on comprehensive video benchmarks. For example, it achieves a 49.7 accuracy on the Temporal Reasoning task of Video-MME, surpassing powerful large-scale models such as InternVL-Chat-V1.5-20B and VILA1.5-40B. Further analysis reveals a strong correlation between textual and video temporal task performance, validating the efficacy of transferring temporal reasoning abilities from text to video domains.

2056. Revealing the 3D Cosmic Web through Gravitationally Constrained Neural Fields

链接：https://iclr.cc/virtual/2025/poster/30603 abstract： Weak gravitational lensing is the slight distortion of galaxy shapes caused primarily by the gravitational effects of dark matter in the universe. In our work, we seek to invert the weak lensing signal from 2D telescope images to reconstruct a 3D map of the universe’s dark matter field. While inversion typically yeilds a 2D projection of the dark matter field, accurate 3D maps of the dark matter distribution are essential for localizing structures of interest and testing theories of our universe. However, 3D inversion poses signficant challenges. First, unlike standard 3D reconstruction that relies on multiple viewpoints, in this case, images are only observed from a single viewpoint. This challenge can be partially addressed by observing how galaxy emitters throughout the volume are lensed. However, this leads to the second challenge: the shapes and exact locations of unlensed galaxies are unknown, and can only be estimated with a very large degree of uncertainty. This introduces an overwhelming amount of noise which nearly drowns out the lensing signal completely. Previous approaches tackle this by imposing strong assumptions about the structures in the volume. We instead propose a methodology using a gravitationally-constrained neural field to flexibly model the continuous matter distribution. We take an analysis-by-synthesis approach, optimizing the weights of the neural network through a fully differentiable physical forward model to reproduce the lensing signal present in image measurements. We showcase our method on simulations, including realistic simulated measurements of dark matter distributions that mimic data from upcoming telescope surveys. Our results show that our method can not only outperform previous methods, but importantly is also able to recover potentially surprising dark matter structures.

2057. Peeking Behind Closed Doors: Risks of LLM Evaluation by Private Data Curators

链接：https://iclr.cc/virtual/2025/poster/31326 abstract： The rapid advancement in building large language models (LLMs) has intensified competition among big-tech companies and AI startups. In this regard, model evaluations are critical for product and investment-related decision-making. While open evaluation sets like MMLU initially drove progress, concerns around data contamination and data bias have constantly questioned their reliability. As a result, it has led to the rise of private data curators who have begun conducting hidden evaluations with high-quality self-curated test prompts and their own expert annotators. In this blog post, we argue that despite potential advantages in addressing contamination issues, private evaluations introduce inadvertent financial and evaluation risks. In particular, the key concerns include the potential conflict of interest arising from private data curators’ business relationships with their clients (leading LLM firms). In addition, we highlight that the subjective preferences of private expert annotators will lead to inherent evaluation bias towards the models trained with the private curators’ data. Overall, this blog post lays the foundation for studying the risks of private evaluations that can lead to wide-ranging community discussions and policy changes.

2058. Reassessing EMNLP 2024’s Best Paper: Does Divergence-Based

Calibration for MIAs Hold Up?

链接：https://iclr.cc/virtual/2025/poster/31327 abstract： At EMNLP 2024, the Best Paper Award was given to "Pretraining Data Detection for Large Language Models: A Divergence-based Calibration Method". The paper addresses Membership Inference Attacks (MIAs), a key issue in machine learning related to privacy. The authors propose a new calibration method and introduce PatentMIA, a benchmark utilizing temporally shifted patent data to validate their approach. The method initially seems promising: it recalibrates model probabilities using a divergence metric between the outputs of a target model and a token-frequency map derived from auxiliary data, claiming improved detection of member and non-member samples. However, upon closer examination, we identified significant shortcomings in both the experimental design and evaluation methodology. In this post, we critically analyze the paper and its broader implications.

2059. Mining your own secrets: Diffusion Classifier Scores for Continual Personalization of Text-to-Image Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/28758 abstract： Personalized text-to-image diffusion models have grown popular for their ability to efficiently acquire a new concept from user-defined text descriptions and a few images. However, in the real world, a user may wish to personalize a model on multiple concepts but one at a time, with no access to the data from previous concepts due to storage/privacy concerns. When faced with this continual learning (CL) setup, most personalization methods fail to find a balance between acquiring new concepts and retaining previous ones -- a challenge that continual personalization (CP) aims to solve. Inspired by the successful CL methods that rely on class-specific information for regularization, we resort to the inherent class-conditioned density estimates, also known as diffusion classifier (DC) scores, for CP of text-to-image diffusion models. Namely, we propose using DC scores for regularizing the parameter-space and function-space of text-to-image diffusion models.Using several diverse evaluation setups, datasets, and metrics, we show that our proposed regularization-based CP methods outperform the state-of-the-art C-LoRA, and other baselines. Finally, by operating in the replay-free CL setup and on low-rank adapters, our method incurs zero storage and parameter overhead, respectively, over the state-of-the-art.

2060. Herald: A Natural Language Annotated Lean 4 Dataset

链接：https://iclr.cc/virtual/2025/poster/29589 abstract： Verifiable formal languages like Lean have profoundly impacted mathematical reasoning, particularly through the use of large language models (LLMs) for automated reasoning. A significant challenge in training LLMs for these formal languages is the lack of parallel datasets that align natural language with formal language proofs. To address this challenge, this paper introduces a novel framework for translating the Mathlib4 corpus (a unified library of mathematics in formal language Lean 4) into natural language. Building upon this, we employ a dual augmentation strategy that combines tactic-based and informal-based approaches, leveraging the Lean-jixia system, a Lean 4 analyzer. We present the results of this pipeline on Mathlib4 as Herald (Hierarchy and Retrieval-based Translated Lean Dataset). We also propose the Herald Translator, which is fine-tuned on Herald. Herald translator achieves a 96.7\% accuracy (Pass@128) on formalizing statements in the miniF2F-test and a 23.5\% accuracy on our internal graduate-level textbook dataset, outperforming InternLM2-Math-Plus-7B (73.0\% and 7.5\%) and TheoremLlama (50.1\% and 4.0\%). Furthermore, we propose a section-level translation framework for real-world applications. As a direct application of Herald translator, we have successfully translated a template section in the Stack project, marking a notable progress in the automatic formalization of graduate-level mathematical literature. Our model, along with the datasets, are open-sourced to the public.

2061. SoftCVI: Contrastive variational inference with self-generated soft labels

链接：https://iclr.cc/virtual/2025/poster/29737 abstract： Estimating a distribution given access to its unnormalized density is pivotal in Bayesian inference, where the posterior is generally known only up to an unknown normalizing constant. Variational inference and Markov chain Monte Carlo methods are the predominant tools for this task; however, both are often challenging to apply reliably, particularly when the posterior has complex geometry. Here, we introduce Soft Contrastive Variational Inference (SoftCVI), which allows a family of variational objectives to be derived through a contrastive estimation framework. The approach parameterizes a classifier in terms of a variational distribution, reframing the inference task as a contrastive estimation problem aiming to identify a single true posterior sample among a set of samples. Despite this framing, we do not require positive or negative samples, but rather learn by sampling the variational distribution and computing ground truth soft classification labels from the unnormalized posterior itself. The objectives have zero variance gradient when the variational approximation is exact, without the need for specialized gradient estimators. We empirically investigate the performance on a variety of Bayesian inference tasks, using both simple (e.g. normal) and expressive (normalizing flow) variational distributions. We find that SoftCVI can be used to form objectives which are stable to train and mass-covering, frequently outperforming inference with other variational approaches.

2062. ContraDiff: Planning Towards High Return States via Contrastive Learning

链接：https://iclr.cc/virtual/2025/poster/29310 abstract： The performance of offline reinforcement learning (RL) is sensitive to the proportion of high-return trajectories in the offline dataset. However, in many simulation environments and real-world scenarios, there are large ratios of low-return trajectories rather than high-return trajectories, which makes learning an efficient policy challenging. In this paper, we propose a method called Contrastive Diffuser (ContraDiff) to make full use of low-return trajectories and improve the performance of offline RL algorithms. Specifically, ContraDiff groups the states of trajectories in the offline dataset into high-return states and low-return states and treats them as positive and negative samples correspondingly. Then, it designs a contrastive mechanism to pull the planned trajectory of an agent toward high-return states and push them away from low-return states. Through the contrast mechanism, trajectories with low returns can serve as negative examples for policy learning, guiding the agent to avoid areas associated with low returns and achieve better performance. Through the contrast mechanism, trajectories with low returns provide a ``counteracting force'' guides the agent to avoid areas associated with low returns and achieve better performance.Experiments on 27 sub-optimal datasets demonstrate the effectiveness of our proposed method. Our code is publicly available at https://github.com/Looomo/contradiff.

2063. Class Distribution-induced Attention Map for Open-vocabulary Semantic Segmentations

链接：https://iclr.cc/virtual/2025/poster/30516 abstract： Open-vocabulary semantic segmentation is a challenging task that assigns seen or unseen class labels to individual pixels. While recent works with vision-language models (VLMs) have shown promising results in zero-shot semantic segmentation, they still struggle to accurately localize class-related objects. In this work, we argue that CLIP-based prior works yield patch-wise noisy class predictions while having highly correlated class distributions for each object. Then, we propose Class Distribution-induced Attention Map, dubbed CDAM, that is generated by the Jensen-Shannon divergence between class distributions of two patches that belong to the same (class) object. This CDAM can be used for open-vocabulary semantic segmentation by integrating it into the final layer of CLIP to enhance the capability to accurately localize desired classes. Our class distribution-induced attention scheme can easily work with multi-scale image patches as well as augmented text prompts for further enhancing attention maps. By exploiting class distribution, we also propose robust entropy-based background thresholding for the inference of semantic segmentation. Interestingly, the core idea of our proposed method does not conflict with other prior arts in zero-shot semantic segmentation, thus can be synergetically used together, yielding substantial improvements in performance across popular semantic segmentation benchmarks.

2064. Concept Pinpoint Eraser for Text-to-image Diffusion Models via Residual Attention Gate

链接：https://iclr.cc/virtual/2025/poster/29221 abstract： Remarkable progress in text-to-image diffusion models has brought a major concern about potentially generating images on inappropriate or trademarked concepts. Concept erasing has been investigated with the goals of deleting target concepts in diffusion models while preserving other concepts with minimal distortion. To achieve these goals, recent concept erasing methods usually fine-tune the cross-attention layers of diffusion models. In this work, we first show that merely updating the cross-attention layers in diffusion models, which is mathematically equivalent to adding linear modules to weights, may not be able to preserve diverse remaining concepts. Then, we propose a novel framework, dubbed Concept Pinpoint Eraser (CPE), by adding nonlinear Residual Attention Gates (ResAGs) that selectively erase (or cut) target concepts while safeguarding remaining concepts from broad distributions by employing an attention anchoring loss to prevent the forgetting. Moreover, we adversarially train CPE with ResAG and learnable text embeddings in an iterative manner to maximize erasing performance and enhance robustness against adversarial attacks. Extensive experiments on the erasure of celebrities, artistic styles, and explicit contents demonstrated that the proposed CPE outperforms prior arts by keeping diverse remaining concepts while deleting the target concepts with robustness against attack prompts. Code is available at https://github.com/Hyun1A/CPE.

2065. Rethinking Fair Representation Learning for Performance-Sensitive Tasks

链接：https://iclr.cc/virtual/2025/poster/28312 abstract： We investigate the prominent class of fair representation learning methods for bias mitigation. Using causal reasoning to define and formalise different sources of dataset bias, we reveal important implicit assumptions inherent to these methods. We prove fundamental limitations on fair representation learning when evaluation data is drawn from the same distribution as training data and run experiments across a range of medical modalities to examine the performance of fair representation learning under distribution shifts. Our results explain apparent contradictions in the existing literature and reveal how rarely considered causal and statistical aspects of the underlying data affect the validity of fair representation learning. We raise doubts about current evaluation practices and the applicability of fair representation learning methods in performance-sensitive settings. We argue that fine-grained analysis of dataset biases should play a key role in the field moving forward.

2066. Reconstruction-Guided Policy: Enhancing Decision-Making through Agent-Wise State Consistency

链接：https://iclr.cc/virtual/2025/poster/29270 abstract： An important challenge in multi-agent reinforcement learning is partial observability, where agents cannot access the global state of the environment during execution and can only receive observations within their field of view. To address this issue, previous works typically use the dimensional-wise state, which is obtained by applying MLP or dimensional-based attention on the global state, for decision-making during training and relying on a reconstructed dimensional-wise state during execution. However, dimensional-wise states tend to divert agent attention to specific features, neglecting potential dependencies between agents, making it difficult to make optimal decisions. Moreover, the inconsistency between the states used in training and execution further increases additional errors. To resolve these issues, we propose a method called Reconstruction-Guided Policy (RGP) to reconstruct the agent-wise state, which represents the information of inter-agent relationships, as input for decision-making during both training and execution. This not only preserves the potential dependencies between agents but also ensures consistency between the states used in training and execution. We conducted extensive experiments on both discrete and continuous action environments to evaluate RGP, and the results demonstrates its superior effectiveness. Our code is public in https://anonymous.4open.science/r/RGP-9F79

2067. BOFormer: Learning to Solve Multi-Objective Bayesian Optimization via Non-Markovian RL

链接：https://iclr.cc/virtual/2025/poster/29456 abstract： Bayesian optimization (BO) offers an efficient pipeline for optimizing black-box functions with the help of a Gaussian process prior and an acquisition function (AF). Recently, in the context of single-objective BO, learning-based AFs witnessed promising empirical results given its favorable non-myopic nature. Despite this, the direct extension of these approaches to multi-objective Bayesian optimization (MOBO) suffer from the hypervolume identifiability issue, which results from the non-Markovian nature of MOBO problems. To tackle this, inspired by the non-Markovian RL literature and the success of Transformers in language modeling, we present a generalized deep Q-learning framework and propose BOFormer, which substantiates this framework for MOBO via sequence modeling. Through extensive evaluation, we demonstrate that BOFormer constantly achieves better performance than the benchmark rule-based and learning-based algorithms in various synthetic MOBO and real-world multi-objective hyperparameter optimization problems.

2068. From Probability to Counterfactuals: the Increasing Complexity of Satisfiability in Pearl's Causal Hierarchy

链接：https://iclr.cc/virtual/2025/poster/28162 abstract： The framework of Pearl's Causal Hierarchy (PCH) formalizes three types of reasoning: probabilistic (i.e. purely observational), interventional, and counterfactual, that reflect the progressive sophistication of human thought regarding causation. We investigate the computational complexity aspects of reasoning in this framework focusing mainly on satisfiability problems expressed in probabilistic and causal languages across the PCH. That is, given a system of formulas in the standard probabilistic and causal languages, does there exist a model satisfying the formulas? Our main contribution is to prove the exact computational complexities showing that languages allowing addition and marginalization (via the summation operator) yield NP^{PP}-, PSPACE-, and NEXP-complete satisfiability problems, depending on the level of the PCH. These are the first results to demonstrate a strictly increasing complexity across the PCH: from probabilistic to causal and counterfactual reasoning. On the other hand, in the case of full languages, i.e.~allowing addition, marginalization, and multiplication, we show that the satisfiability for the counterfactual level remains the same as for the probabilistic and causal levels, solving an open problem in the field.

2069. Exploring The Loss Landscape Of Regularized Neural Networks Via Convex Duality

链接：https://iclr.cc/virtual/2025/poster/30981 abstract： We discuss several aspects of the loss landscape of regularized neural networks: the structure of stationary points, connectivity of optimal solutions, path with non-increasing loss to arbitrary global optimum, and the nonuniqueness of optimal solutions, by casting the problem into an equivalent convex problem and considering its dual. Starting from two-layer neural networks with scalar output, we first characterize the solution set of the convex problem using its dual and further characterize all stationary points. With the characterization, we show that the topology of the global optima goes through a phase transition as the width of the network changes, and construct counterexamples where the problem may have a continuum of optimal solutions. Finally, we show that the solution set characterization and connectivity results can be extended to different architectures, including two layer vector-valued neural networks and parallel three-layer neural networks.

2070. Reasoning Elicitation in Language Models via Counterfactual Feedback

链接：https://iclr.cc/virtual/2025/poster/29410 abstract： Despite the increasing effectiveness of language models, their reasoning capabilities remain underdeveloped. In particular, causal reasoning through counterfactual question answering is lacking. This work aims to bridge this gap. We first derive novel metrics that balance accuracy in factual and counterfactual questions, capturing a more complete view of the reasoning abilities of language models than traditional factual-only based metrics. Second, we propose several fine-tuning approaches that aim to elicit better reasoning mechanisms, in the sense of the proposed metrics. Finally, we evaluate the performance of the fine-tuned language models in a variety of realistic scenarios. In particular, we investigate to what extent our fine-tuning approaches systemically achieve better generalization with respect to the base models in several problems that require, among others, inductive and deductive reasoning capabilities.

2071. TSVD: Bridging Theory and Practice in Continual Learning with Pre-

trained Models

链接：https://iclr.cc/virtual/2025/poster/29087 abstract： The goal of continual learning (CL) is to train a model that can solve multipletasks presented sequentially. Recent CL approaches have achieved strong performanceby leveraging large pre-trained models that generalize well to downstreamtasks. However, such methods lack theoretical guarantees, making them prone tounexpected failures. Conversely, principled CL approaches often fail to achievecompetitive performance. In this work, we aim to bridge this gap between theoryand practice by designing a simple CL method that is theoretically sound andhighly performant. Specifically, we lift pre-trained features into a higher dimensionalspace and formulate an over-parametrized minimum-norm least-squaresproblem. We find that the lifted features are highly ill-conditioned, potentiallyleading to large training errors (numerical instability) and increased generalizationerrors. We address these challenges by continually truncating the singular valuedecomposition (SVD) of the lifted features. Our approach, termed TSVD, is stablewith respect to the choice of hyperparameters, can handle hundreds of tasks, andoutperforms state-of-the-art CL methods on multiple datasets. Importantly, ourmethod satisfies a recurrence relation throughout its continual learning process,which allows us to prove it maintains small training and generalization errors byappropriately truncating a fraction of SVD factors. This results in a stable continuallearning method with strong empirical performance and theoretical guarantees.Code available: https://github.com/liangzu/tsvd.

2072. InCoDe: Interpretable Compressed Descriptions For Image Generation

链接：https://iclr.cc/virtual/2025/poster/32075 abstract： Generative models have been successfully applied in diverse domains, from natural language processing to image synthesis. However, despite this success, a key challenge that remains is the ability to control the semantic content of the scene being generated. We argue that adequate control of the generation process requires a data representation that allows users to access and efficiently manipulate the semantic factors shaping the data distribution. This work advocates for the adoption of succinct, informative, and interpretable representations, quantified using information-theoretic principles. Through extensive experiments, we demonstrate the efficacy of our proposed framework both qualitatively and quantitatively. Our work contributes to the ongoing quest to enhance both controllability and interpretability in the generation process. Code available at github.com/ArmandCom/InCoDe.

2073. Coreset Spectral Clustering

链接：https://iclr.cc/virtual/2025/poster/31180 abstract： Coresets have become an invaluable tool for solving $k$-means and kernel $k$-means clustering problems on large datasets with small numbers of clusters. On the other hand, spectral clustering works well on sparse graphs and has recently been extended to scale efficiently to large numbers of clusters. We exploit the connection between kernel $k$-means and the normalised cut problem to combine the benefits of both. Our main result is a coreset spectral clustering algorithm for graphs that clusters a coreset graph to infer a good labelling of the original graph. We prove that an $\alpha$-approximation for the normalised cut problem on the coreset graph is an $O(\alpha)$-approximation on the original. We also improve the running time of the state-of-the-art coreset algorithm for kernel $k$-means on sparse kernels, from $\tilde{O}(nk)$ to $\tilde{O}(n\cdot \min (k, d_{avg}))$, where $d_{avg}$ is the average number of non-zero entries in each row of the $n\times n$ kernel matrix. Our experiments confirm our coreset algorithm is asymptotically faster on large real-world graphs with many clusters, and show that our clustering algorithm overcomes the main challenge faced by coreset kernel $k$-means on sparse kernels which is getting stuck in local optima.

2074. ObscuraCoder: Powering Efficient Code LM Pre-Training Via Obfuscation Grounding

链接：https://iclr.cc/virtual/2025/poster/29407 abstract： Language models (LMs) have become a staple of the code-writing toolbox. Their pre-training recipe has, however, remained stagnant over recent years, barring the occasional changes in data sourcing and filtering strategies. In particular, research exploring modifications to Code-LMs' pre-training objectives, geared towards improving data efficiency and better disentangling between syntax and semantics, has been noticeably sparse, especially compared with corresponding efforts in natural language LMs. In this work, we examine grounding on obfuscated code as a means of helping Code-LMs look beyond the surface-form syntax and enhance their pre-training sample efficiency. To this end, we compile ObscuraX, a dataset of approximately 55M source and obfuscated code pairs in seven languages. Subsequently, we pre-train ObscuraCoder models, ranging in size from 255M to 2.8B parameters, on a 272B-token corpus that includes ObscuraX and demonstrate that our obfuscation-based pre-training recipe leads to consistent improvements in Code-LMs' abilities compared to both vanilla autoregressive pre-training as well as existing de-obfuscation (DOBF) objectives. ObscuraCoder demonstrates sizeable gains across multiple tests of syntactic and semantic code understanding, along with improved capabilities in multilingual code completion, multilingual code commit summarization, and multi-purpose library-oriented code generation.

2075. Diffusion Attribution Score: Evaluating Training Data Influence in Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/28556 abstract： As diffusion models become increasingly popular, the misuse of copyrighted and private images has emerged as a major concern. One promising solution to mitigate this issue is identifying the contribution of specific training samples in generative models, a process known as data attribution. Existing data attribution methods for diffusion models typically quantify the contribution of a training sample by evaluating the change in diffusion loss when the sample is included or excluded from the training process.However, we argue that the direct usage of diffusion loss cannot represent such a contribution accurately due to the calculation of diffusion loss.Specifically, these approaches measure the divergence between predicted and ground truth distributions, which leads to an indirect comparison between the predicted distributions and cannot represent the variances between model behaviors.To address these issues, we aim to measure the direct comparison between predicted distributions with an attribution score to analyse the training sample importance, which is achieved by Diffusion Attribution Score (\textit{DAS}).Underpinned by rigorous theoretical analysis, we elucidate the effectiveness of DAS.Additionally, we explore strategies to accelerate DAS calculations, facilitating its application to large-scale diffusion models.Our extensive experiments across various datasets and diffusion models demonstrate that DAS significantly surpasses previous benchmarks in terms of the linear data-modelling score, establishing new state-of-the-art performance.

2076. Efficient and Accurate Explanation Estimation with Distribution Compression

链接：https://iclr.cc/virtual/2025/poster/29979 abstract： We discover a theoretical connection between explanation estimation and distribution compression that significantly improves the approximation of feature attributions, importance, and effects. While the exact computation of various machine learning explanations requires numerous model inferences and becomes impractical, the computational cost of approximation increases with an ever-increasing size of data and model parameters. We show that the standard i.i.d. sampling used in a broad spectrum of algorithms for post-hoc explanation leads to an approximation error worthy of improvement. To this end, we introduce Compress Then Explain (CTE), a new paradigm of sample-efficient explainability. It relies on distribution compression through kernel thinning to obtain a data sample that best approximates its marginal distribution. CTE significantly improves the accuracy and stability of explanation estimation with negligible computational overhead. It often achieves an on-par explanation approximation error 2-3x faster by using fewer samples, i.e. requiring 2-3x fewer model evaluations. CTE is a simple, yet powerful, plug-in for any explanation method that now relies on i.i.d. sampling.

2077. Fundamental Limitations on Subquadratic Alternatives to Transformers

链接：https://iclr.cc/virtual/2025/poster/29565 abstract： The Transformer architecture is widely deployed in many popular and impactful Large Language Models. At its core is the attention mechanism for calculating correlations between pairs of tokens. Performing an attention computation takes quadratic time in the input size, and had become the time bottleneck for transformer operations. In order to circumvent this, researchers have used a variety of approaches, including designing heuristic algorithms for performing attention computations faster, and proposing alternatives to the attention mechanism which can be computed more quickly. For instance, state space models such as Mamba were designed to replace attention with an almost linear time alternative.In this paper, we prove that any such approach cannot perform important tasks that Transformer is able to perform (assuming a popular conjecture from fine-grained complexity theory). We focus on document similarity tasks, where one is given as input many documents and would like to find a pair which is (approximately) the most similar. We prove that Transformer is able to perform this task, and we prove that this task cannot be performed in truly subquadratic time by any algorithm. Thus, any model which can be evaluated in subquadratic time – whether because of subquadratic-time heuristics for attention, faster attention replacements like Mamba, or any other reason – cannot perform this task. In other words, in order to perform tasks that (implicitly or explicitly) involve document similarity, one may as well use Transformer and cannot avoid its quadratic running time.

2078. Latent-EnSF: A Latent Ensemble Score Filter for High-Dimensional Data Assimilation with Sparse Observation Data

链接：https://iclr.cc/virtual/2025/poster/27953 abstract： Accurate modeling and prediction of complex physical systems often rely on data assimilation techniques to correct errors inherent in model simulations. Traditional methods like the Ensemble Kalman Filter (EnKF) and its variants as well as the recently developed Ensemble Score Filters (EnSF) face significant challenges when dealing with high-dimensional and nonlinear Bayesian filtering problems with sparse observations, which are ubiquitous in real-world applications. In this paper, we propose a novel data assimilation method, Latent-EnSF, which leverages EnSF with efficient and consistent latent representations of the full states and sparse observations to address the joint challenges of high dimensionlity in states and high sparsity in observations for nonlinear Bayesian filtering. We introduce a coupled Variational Autoencoder (VAE) with two encoders to encode the full states and sparse observations in a consistent way guaranteed by a latent distribution matching and regularization as well as a consistent state reconstruction. With comparison to several methods, we demonstrate the higher accuracy, faster convergence, and higher efficiency of Latent-EnSF for two challenging applications with complex models in shallow water wave propagation and medium-range weather forecasting, for highly sparse observations in both space and time.

2079. A Meta-Learning Approach to Bayesian Causal Discovery

链接：https://iclr.cc/virtual/2025/poster/28913 abstract： Discovering a unique causal structure is difficult due to both inherent identifiability issues, and the consequences of finite data.As such, uncertainty over causal structures, such as those obtained from a Bayesian posterior, are often necessary for downstream tasks.Finding an accurate approximation to this posterior is challenging, due to the large number of possible causal graphs, as well as the difficulty in the subproblem of finding posteriors over the functional relationships of the causal edges.Recent works have used Bayesian meta learning to view the problem of posterior estimation as a supervised learning task.Yet, these methods are limited as they cannot reliably sample from the posterior over causal structures and fail to encode key properties of the posterior, such as correlation between edges and permutation equivariance with respect to nodes.To address these limitations, we propose a Bayesian meta learning model that allows for sampling causal structures from the posterior and encodes these key properties.We compare our meta-Bayesian causal discovery against existing Bayesian causal discovery methods, demonstrating the advantages of directly learning a posterior over causal structure.

2080. Start Smart: Leveraging Gradients For Enhancing Mask-based XAI Methods

链接：https://iclr.cc/virtual/2025/poster/30158 abstract： Mask-based explanation methods offer a powerful framework for interpreting deep learning model predictions across diverse data modalities, such as images and time series, in which the central idea is to identify an instance-dependent mask that minimizes the performance drop from the resulting masked input. Different objectives for learning such masks have been proposed, all of which, in our view, can be unified under an information-theoretic framework that balances performance degradation of the masked input with the complexity of the resulting masked representation. Typically, these methods initialize the masks either uniformly or as all-ones.In this paper, we argue that an effective mask initialization strategy is as important as the development of novel learning objectives, particularly in light of the significant computational costs associated with existing mask-based explanation methods. To this end, we introduce a new gradient-based initialization technique called StartGrad, which is the first initialization method specifically designed for mask-based post-hoc explainability methods. Compared to commonly used strategies, StartGrad is provably superior at initialization in striking the aforementioned trade-off. Despite its simplicity, our experiments demonstrate that StartGrad enhances the optimization process of various state-of-the-art mask-explanation methods by reaching target metrics faster and, in some cases, boosting their overall performance.

2081. Learning Chaos In A Linear Way

链接：https://iclr.cc/virtual/2025/poster/29977 abstract： Learning long-term behaviors in chaotic dynamical systems, such as turbulent flows and climate modelling, is challenging due to their inherent instability and unpredictability. These systems exhibit positive Lyapunov exponents, which significantly hinder accurate long-term forecasting. As a result, understanding long-term statistical behavior is far more valuable than focusing on short-term accuracy. While autoregressive deep sequence models have been applied to capture long-term behavior, they often lead to exponentially increasing errors in learned dynamics. To address this, we shift the focus from simple prediction errors to preserving an invariant measure in dissipative chaotic systems. These systems have attractors, where trajectories settle, and the invariant measure is the probability distribution on attractors that remains unchanged under dynamics. Existing methods generate long trajectories of dissipative chaotic systems by aligning invariant measures, but it is not always possible to obtain invariant measures for arbitrary datasets. We propose the Poincaré Flow Neural Network (PFNN), a novel operator learning framework designed to capture behaviors of chaotic systems without any explicit knowledge of the invariant measure. PFNN employs an auto-encoder to map the chaotic system to a finite-dimensional feature space, effectively linearizing the chaotic evolution. It then learns the linear evolution operators to match the physical dynamics by addressing two critical properties in dissipative chaotic systems: (1) contraction, the system’s convergence toward its attractors, and (2) measure invariance, trajectories on the attractors following a probability distribution invariant to the dynamics. Our experiments on a variety of chaotic systems, including Lorenz systems, Kuramoto-Sivashinsky equation and Navier–Stokes equation, demonstrate that PFNN has more accurate predictions and physical statistics compared to competitive baselines including the Fourier Neural Operator and the Markov Neural Operator.

2082. Improving Instruction-Following in Language Models through Activation Steering

链接：https://iclr.cc/virtual/2025/poster/27819 abstract： The ability to follow instructions is crucial for numerous real-world applications of language models. In pursuit of deeper insights and more powerful capabilities, we derive instruction-specific vector representations from language models and use them to steer models accordingly. These vectors are computed as the difference in activations between inputs with and without instructions, enabling a modular approach to activation steering. We demonstrate how this method can enhance model adherence to constraints such as output format, length, and word inclusion, providing inference-time control over instruction following. Our experiments across four models demonstrate how we can use the activation vectors to guide models to follow constraints even without explicit instructions and to enhance performance when instructions are present. Additionally, we explore the compositionality of activation steering, successfully applying multiple instructions simultaneously. Finally, we demonstrate that steering vectors computed on instruction-tuned models can transfer to improve base models. Our findings demonstrate that activation steering offers a practical and scalable approach for fine-grained control in language generation. Our code and data are available at https://github.com/microsoft/llm-steer-instruct.

2083. Unearthing Skill-level Insights for Understanding Trade-offs of Foundation Models

链接：https://iclr.cc/virtual/2025/poster/28591 abstract： With models getting stronger, evaluations have grown more complex, testing multiple skills in one benchmark and even in the same instance at once. However, skill-wise performance is obscured when inspecting aggregate accuracy, under-utilizing the rich signal modern benchmarks contain. We propose an automatic approach to recover the underlying skills relevant for any evaluation instance, by way of inspecting model-generated {\em rationales}. After validating the relevance of rationale-parsed skills and inferring skills for 46k instances over 12 benchmarks, we observe many skills to be common across benchmarks, resulting in the curation of hundreds of \emph{skill-slices} (i.e. sets of instances testing a common skill). Inspecting accuracy over these slices yields novel insights on model trade-offs: e.g., compared to GPT-4o and Claude 3.5 Sonnet, on average, Gemini 1.5 Pro is 18% more accurate in \emph{computing molar mass}, but 19% less accurate in \emph{applying constitutional law}, despite the overall accuracies of the three models differing by a mere 0.4%. Furthermore, we demonstrate the practical utility of our approach by showing that insights derived from skill slice analysis can generalize to held-out instances: when routing each instance to the model strongest on the relevant skills, we see a 3% accuracy improvement over our 12 dataset corpus. Our skill-slices and framework open a new avenue in model evaluation, leveraging skill-specific analyses to unlock a more granular and actionable understanding of model capabilities.

2084. Union-over-Intersections: Object Detection beyond Winner-Takes-All

链接：https://iclr.cc/virtual/2025/poster/30208 abstract： This paper revisits the problem of predicting box locations in object detection architectures. Typically, each box proposal or box query aims to directly maximize the intersection-over-union score with the ground truth, followed by a winner-takes-all non-maximum suppression where only the highest scoring box in each region is retained. We observe that both steps are sub-optimal: the first involves regressing proposals to the entire ground truth, which is a difficult task even with large receptive fields, and the second neglects valuable information from boxes other than the top candidate. Instead of regressing proposals to the whole ground truth, we propose a simpler approach—regress only to the area of intersection between the proposal and the ground truth. This avoids the need for proposals to extrapolate beyond their visual scope, improving localization accuracy. Rather than adopting a winner-takes-all strategy, we take the union over the regressed intersections of all boxes in a region to generate the final box outputs. Our plug-and-play method integrates seamlessly into proposal-based, grid-based, and query-based detection architectures with minimal modifications, consistently improving object localization and instance segmentation. We demonstrate its broad applicability and versatility across various detection and segmentation tasks.

2085. Compositional Entailment Learning for Hyperbolic Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/31060 abstract： Image-text representation learning forms a cornerstone in vision-language models, where pairs of images and textual descriptions are contrastively aligned in a shared embedding space. Since visual and textual concepts are naturally hierarchical, recent work has shown that hyperbolic space can serve as a high-potential manifold to learn vision-language representation with strong downstream performance. In this work, for the first time we show how to fully leverage the innate hierarchical nature of hyperbolic embeddings by looking beyond individual image-text pairs. We propose Compositional Entailment Learning for hyperbolic vision-language models. The idea is that an image is not only described by a sentence but is itself a composition of multiple object boxes, each with their own textual description. Such information can be obtained freely by extracting nouns from sentences and using openly available localized grounding models. We show how to hierarchically organize images, image boxes, and their textual descriptions through contrastive and entailment-based objectives. Empirical evaluation on a hyperbolic vision-language model trained with millions of image-text pairs shows that the proposed compositional learning approach outperforms conventional Euclidean CLIP learning, as well as recent hyperbolic alternatives, with better zero-shot and retrieval generalization and clearly stronger hierarchical performance.

2086. Implicit Neural Surface Deformation with Explicit Velocity Fields

链接：https://iclr.cc/virtual/2025/poster/28123 abstract： In this work, we introduce the first unsupervised method that simultaneously predicts time-varying neural implicit surfaces and deformations between pairs of point clouds. We propose to model the point movement using an explicit velocity field and directly deform a time-varying implicit field using the modified level-set equation. This equation utilizes an iso-surface evolution with Eikonal constraints in a compact formulation, ensuring the integrity of the signed distance field. By applying a smooth, volume-preserving constraint to the velocity field, our method successfully recovers physically plausible intermediate shapes. Our method is able to handle both rigid and non-rigid deformations without any intermediate shape supervision. Our experimental results demonstrate that our method significantly outperforms existing works, delivering superior results in both quality and efficiency.

2087. Intelligence at the Edge of Chaos

链接：https://iclr.cc/virtual/2025/poster/30160 abstract： We explore the emergence of intelligent behavior in artificial systems by investigating how the complexity of rule-based systems influences the capabilities of models trained to predict these rules. Our study focuses on elementary cellular automata (ECA), simple yet powerful one-dimensional systems that generate behaviors ranging from trivial to highly complex. By training distinct Large Language Models (LLMs) on different ECAs, we evaluated the relationship between the complexity of the rules' behavior and the intelligence exhibited by the LLMs, as reflected in their performance on downstream tasks. Our findings reveal that rules with higher complexity lead to models exhibiting greater intelligence, as demonstrated by their performance on reasoning and chess move prediction tasks. Both uniform and periodic systems, and often also highly chaotic systems, resulted in poorer downstream performance, highlighting a sweet spot of complexity conducive to intelligence. We conjecture that intelligence arises from the ability to predict complexity and that creating intelligence may require only exposure to complexity.

2088. Multimodal Situational Safety

链接：https://iclr.cc/virtual/2025/poster/30187 abstract： Multimodal Large Language Models (MLLMs) are rapidly evolving, demonstrating impressive capabilities as multimodal assistants that interact with both humans and their environments. However, this increased sophistication introduces significant safety concerns. In this paper, we present the first evaluation and analysis of a novel safety challenge termed Multimodal Situational Safety, which explores how safety considerations vary based on the specific situation in which the user or agent is engaged. We argue that for an MLLM to respond safely—whether through language or action—it often needs to assess the safety implications of a language query within its corresponding visual context.To evaluate this capability, we develop the Multimodal Situational Safety benchmark (MSSBench) to assess the situational safety performance of current MLLMs. The dataset comprises 1,960 language query-image pairs, half of which the image context is safe, and the other half is unsafe. We also develop an evaluation framework that analyzes key safety aspects, including explicit safety reasoning, visual understanding, and, crucially, situational safety reasoning. Our findings reveal that current MLLMs struggle with this nuanced safety problem in the instruction-following setting and struggle to tackle these situational safety challenges all at once, highlighting a key area for future research. Furthermore, we develop multi-agent pipelines to coordinately solve safety challenges, which shows consistent improvement in safety over the original MLLM response.

2089. Analysing The Spectral Biases in Generative Models

链接：https://iclr.cc/virtual/2025/poster/31344 abstract： Diffusion and GAN models have demonstrated remarkable success in synthesizing high-quality images propelling them into various real-life applications across different domains. However, it has been observed that they exhibit spectral biases that impact their ability to generate certain frequencies and makes it pretty straightforward to distinguish real images from fake ones. In this blog we analyze these models and attempt to explain the reason behind these biases.

2090. Learning to Communicate Through Implicit Communication Channels

链接：https://iclr.cc/virtual/2025/poster/27824 abstract： Effective communication is an essential component in collaborative multi-agent systems. Situations where explicit messaging is not feasible have been common in human society throughout history, which motivate the study of implicit communication. Previous works on learning implicit communication mostly rely on theory of mind (ToM), where agents infer the mental states and intentions of others by interpreting their actions. However, ToM-based methods become less effective in making accurate inferences in complex tasks. In this work, we propose the Implicit Channel Protocol (ICP) framework, which allows agents to communicate through implicit communication channels similar to the explicit ones. ICP leverages a subset of actions, denoted as the scouting actions, and a mapping between information and these scouting actions that encodes and decodes the messages. We propose training algorithms for agents to message and act, including learning with a randomly initialized information map and with a delayed information map. The efficacy of ICP has been tested on the tasks of Guessing Numbers, Revealing Goals, and Hanabi, where ICP significantly outperforms baseline methods through more efficient information transmission.

2091. Discrete GCBF Proximal Policy Optimization for Multi-agent Safe Optimal Control

链接：https://iclr.cc/virtual/2025/poster/31197 abstract： Control policies that can achieve high task performance and satisfy safety constraints are desirable for any system, including multi-agent systems (MAS). One promising technique for ensuring the safety of MAS is distributed control barrier functions (CBF). However, it is difficult to design distributed CBF-based policies for MAS that can tackle unknown discrete-time dynamics, partial observability, changing neighborhoods, and input constraints, especially when a distributed high-performance nominal policy that can achieve the task is unavailable. To tackle these challenges, we propose DGPPO, a new framework that simultaneously learns both a discrete graph CBF which handles neighborhood changes and input constraints, and a distributed high-performance safe policy for MAS with unknown discrete-time dynamics.We empirically validate our claims on a suite of multi-agent tasks spanning three different simulation engines. The results suggest that, compared with existing methods, our DGPPO framework obtains policies that achieve high task performance (matching baselines that ignore the safety constraints), and high safety rates (matching the most conservative baselines), with a constant set of hyperparameters across all environments.

2092. Learning Continually by Spectral Regularization

链接：https://iclr.cc/virtual/2025/poster/30217 abstract： Loss of plasticity is a phenomenon where neural networks can become more difficult to train over the course of learning. Continual learning algorithms seek to mitigate this effect by sustaining good performance while maintaining network trainability. We develop a new technique for improving continual learning inspired by the observation that the singular values of the neural network parameters at initialization are an important factor for trainability during early phases of learning. From this perspective, we derive a new spectral regularizer for continual learning that better sustains these beneficial initialization properties throughout training. In particular, the regularizer keeps the maximum singular value of each layer close to one. Spectral regularization directly ensures that gradient diversity is maintained throughout training, which promotes continual trainability, while minimally interfering with performance in a single task. We present an experimental analysis that shows how the proposed spectral regularizer can sustain trainability and performance across a range of model architectures in continual supervised and reinforcement learning settings. Spectral regularization is less sensitive to hyperparameters while demonstrating better training in individual tasks, sustaining trainability as new tasks arrive, and achieving better generalization performance..

2093. MGDA Converges under Generalized Smoothness, Provably

链接：https://iclr.cc/virtual/2025/poster/27830 abstract： Multi-objective optimization (MOO) is receiving more attention in various fields such as multi-task learning. Recent works provide some effective algorithms with theoretical analysis but they are limited by the standard L-smooth or bounded-gradient assumptions, which typically do not hold for neural networks, such as Long short-term memory (LSTM) models and Transformers. In this paper, we study a more general and realistic class of generalized ℓ-smooth loss functions, where ℓ is a general non-decreasing function of gradient norm. We revisit and analyze the fundamental multiple gradient descent algorithm (MGDA) and its stochastic version with double sampling for solving the generalized ℓ-smooth MOO problems, which approximate the conflict-avoidant (CA) direction that maximizes the minimum improvement among objectives. We provide a comprehensive convergence analysis of these algorithms and show that they converge to an ϵ-accurate Pareto stationary point with a guaranteed ϵ-level average CA distance (i.e., the gap between the updating direction and the CA direction) over all iterations, where totally $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-4})$ samples are needed for deterministic and stochastic settings, respectively. We prove that they can also guarantee a tighter ϵ-level CA distance in each iteration using more samples. Moreover, we analyze an efficient variant of MGDA named MGDA-FA using only $\mathcal{O}(1)$ time and space, while achieving the same performance guarantee as MGDA.

2094. Boundary constrained Gaussian processes for robust physics-informed machine learning of linear partial differential equations

链接：https://iclr.cc/virtual/2025/poster/31382 abstract： We introduce a framework for designing boundary constrained Gaussian process (BCGP) priors for exact enforcement of linear boundary conditions, and apply it to the machine learning of (initial) boundary value problems involving linear partial differential equations (PDEs).In contrast to existing work, we illustrate how to design boundary constrained mean and kernel functions for all classes of boundary conditions typically used in PDE modelling, namely Dirichlet, Neumann, Robin and mixed conditions. Importantly, this is done in a manner which allows for both forward and inverse problems to be naturally accommodated. We prove that the BCGP kernel has a universal representational capacity under Dirichlet conditions, and establish a formal equivalence between BCGPs and boundary-constrained neural networks (BCNNs) of infinite width.Finally, extensive numerical experiments are performed involving several linear PDEs, the results of which demonstrate the effectiveness and robustness of BCGP inference in the presence of sparse, noisy data.

2095. Multi-Task Dense Predictions via Unleashing the Power of Diffusion

链接：https://iclr.cc/virtual/2025/poster/29497 abstract： Diffusion models have exhibited extraordinary performance in dense prediction tasks. However, there are few works exploring the diffusion pipeline for multi-task dense predictions. In this paper, we unlock the potential of diffusion models in solving multi-task dense predictions and propose a novel diffusion-based method, called TaskDiffusion, which leverages the conditional diffusion process in the decoder. Instead of denoising the noisy labels for different tasks separately, we propose a novel joint denoising diffusion process to capture the task relations during denoising. To be specific, our method first encodes the task-specific labels into a task-integration feature space to unify the encoding strategy. This allows us to get rid of the cumbersome task-specific encoding process. In addition, we also propose a cross-task diffusion decoder conditioned on task-specific multi-level features, which can model the interactions among different tasks and levels explicitly while preserving efficiency. Experiments show that our TaskDiffusion outperforms previous state-of-the-art methods for all dense prediction tasks on the widely-used PASCAL-Context and NYUD-v2 datasets. Our code is available at https://github.com/YuqiYang213/TaskDiffusion.

2096. Law of the Weakest Link: Cross Capabilities of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29519 abstract： The development and evaluation of Large Language Models (LLMs) have largely focused on individual capabilities. However, this overlooks the intersection of multiple abilities across different types of expertise that are often required for real-world tasks, which we term cross capabilities. To systematically explore this concept, we first define seven core individual capabilities and then pair them to form seven common cross capabilities, each supported by a manually constructed taxonomy. Building on these definitions, we introduce CrossEval, a benchmark comprising 1,400 human-annotated prompts, with 100 prompts for each individual and cross capability. To ensure reliable evaluation, we involve expert annotators to assess 4,200 model responses, gathering 8,400 human ratings with detailed explanations to serve as reference examples. Our findings reveal that current LLMs consistently exhibit the ``Law of the Weakest Link,'' where cross-capability performance is significantly constrained by the weakest component. Across 58 cross-capability scores from 17 models, 38 scores are lower than all individual capabilities, while 20 fall between strong and weak, but closer to the weaker ability. These results highlight LLMs' underperformance in cross-capability tasks, emphasizing the need to identify and improve their weakest capabilities as a key research priority. The code, benchmarks, and evaluations are available on our project website.

2097. InvestESG: A multi-agent reinforcement learning benchmark for studying climate investment as a social dilemma

链接：https://iclr.cc/virtual/2025/poster/31135 abstract： InvestESG is a novel multi-agent reinforcement learning (MARL) benchmark designed to study the impact of Environmental, Social, and Governance (ESG) disclosure mandates on corporate climate investments. The benchmark models an intertemporal social dilemma where companies balance short-term profit losses from climate mitigation efforts and long-term benefits from reducing climate risk, while ESG-conscious investors attempt to influence corporate behavior through their investment decisions. Companies allocate capital across mitigation, greenwashing, and resilience, with varying strategies influencing climate outcomes and investor preferences. We are releasing open-source versions of InvestESG in both PyTorch and JAX, which enable scalable and hardware-accelerated simulations for investigating competing incentives in mitigate climate change. Our experiments show that without ESG-conscious investors with sufficient capital, corporate mitigation efforts remain limited under the disclosure mandate. However, when a critical mass of investors prioritizes ESG, corporate cooperation increases, which in turn reduces climate risks and enhances long-term financial stability. Additionally, providing more information about global climate risks encourages companies to invest more in mitigation, even without investor involvement. Our findings align with empirical research using real-world data, highlighting MARL's potential to inform policy by providing insights into large-scale socio-economic challenges through efficient testing of alternative policy and market designs.

2098. Bayesian Regularization of Latent Representation

链接：https://iclr.cc/virtual/2025/poster/29414 abstract： The effectiveness of statistical and machine learning methods depends on how well data features are characterized. Developing informative and interpretable latent representations with controlled complexity is essential for visualizing data structure and for facilitating efficient model building through dimensionality reduction. Latent variable models, such as Gaussian Process Latent Variable Models (GP-LVM), have become popular for learning complex, nonlinear representations as alternatives to Principal Component Analysis (PCA). In this paper, we propose a novel class of latent variable models based on the recently introduced Q-exponential process (QEP), which generalizes GP-LVM with a tunable complexity parameter, $q>0$. Our approach, the \emph{Q-exponential Process Latent Variable Model (QEP-LVM)}, subsumes GP-LVM as a special case when $q=2$, offering greater flexibility in managing representation complexity while enhancing interpretability. To ensure scalability, we incorporate sparse variational inference within a Bayesian training framework. We establish connections between QEP-LVM and probabilistic PCA, demonstrating its superior performance through experiments on datasets such as the Swiss roll, oil flow, and handwritten digits.

2099. Boosting Ray Search Procedure of Hard-label Attacks with Transfer-based Priors

链接：https://iclr.cc/virtual/2025/poster/28063 abstract： One of the most practical and challenging types of black-box adversarial attacks is the hard-label attack, where only top-1 predicted labels are available. One effective approach is to search for the optimal ray direction from the benign image that minimizes the ℓ_p norm distance to the adversarial region. The unique advantage of this approach is that it transforms the hard-label attack into a continuous optimization problem. The objective function value is the ray's radius and can be obtained through a binary search with high query cost. Existing methods use a "sign trick" in gradient estimation to reduce queries. In this paper, we theoretically analyze the quality of this gradient estimation, proposing a novel prior-guided approach to improve ray search efficiency, based on theoretical and experimental analysis. Specifically, we utilize the transfer-based priors from surrogate models, and our gradient estimators appropriately integrate them by approximating the projection of the true gradient onto the subspace spanned by these priors and some random directions, in a query-efficient way. We theoretically derive the expected cosine similarity between the obtained gradient estimators and the true gradient, and demonstrate the improvement brought by using priors. Extensive experiments on the ImageNet and CIFAR-10 datasets show that our approach significantly outperforms 11 state-of-the-art methods in terms of query efficiency.

2100. BOND: Aligning LLMs with Best-of-N Distillation

链接：https://iclr.cc/virtual/2025/poster/31233 abstract： Reinforcement learning from human feedback (RLHF) is a key driver of quality and safety in state-of-the-art large language models.Yet, a surprisingly simple and strong inference-time strategy is Best-of-N sampling that selects the best generation among N candidates.In this paper, we propose Best-of-N Distillation (BOND), a novel RLHF algorithm that seeks to emulate Best-of-N but without its significant computational overhead at inference time. Specifically, BOND is a distribution matching algorithm that forces the distribution of generations from the policy to get closer to the Best-of-N distribution. We use the Jeffreys divergence (a linear combination of forward and backward KL) to balance between mode-covering and mode-seeking behavior, and derive an iterative formulation that utilizes a moving anchor for efficiency. We demonstrate the effectiveness of our approach and several design choices through experiments on abstractive summarization and Gemma models.

2101. Can LLMs Solve Longer Math Word Problems Better?

链接：https://iclr.cc/virtual/2025/poster/30529 abstract： Math Word Problems (MWPs) play a vital role in assessing the capabilities of Large Language Models (LLMs), yet current research primarily focuses on questions with concise contexts. The impact of longer contexts on mathematical reasoning remains under-explored. This study pioneers the investigation of Context Length Generalizability (CoLeG), which refers to the ability of LLMs to solve MWPs with extended narratives. We introduce Extended Grade-School Math (E-GSM), a collection of MWPs featuring lengthy narratives, and propose two novel metrics to evaluate the efficacy and resilience of LLMs in tackling these problems. Our analysis of existing zero-shot prompting techniques with proprietary LLMs along with open-source LLMs reveals a general deficiency in CoLeG. To alleviate these issues, we propose tailored approaches for different categories of LLMs. For proprietary LLMs, we introduce a new instructional prompt designed to mitigate the impact of long contexts. For open-source LLMs, we develop a novel auxiliary task for fine-tuning to enhance CoLeG. Our comprehensive results demonstrate the effectiveness of our proposed methods, showing improved performance on E-GSM. Additionally, we conduct an in-depth analysis to differentiate the effects of semantic understanding and reasoning efficacy, showing that our methods improves the latter. We also establish the generalizability of our methods across several other MWP benchmarks. Our findings highlight the limitations of current LLMs and offer practical solutions correspondingly, paving the way for further exploration of model generalizability and training methodologies.

2102. UGMathBench: A Diverse and Dynamic Benchmark for Undergraduate-Level Mathematical Reasoning with Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28857 abstract： Large Language Models (LLMs) have made significant strides in mathematical reasoning, underscoring the need for a comprehensive and fair evaluation of their capabilities. However, existing benchmarks often fall short, either lacking extensive coverage of undergraduate-level mathematical problems or probably suffering from test-set contamination. To address these issues, we introduce UGMathBench, a diverse and dynamic benchmark specifically designed for evaluating undergraduate-level mathematical reasoning with LLMs. UGMathBench comprises 5,062 problems across 16 subjects and 111 topics, featuring 10 distinct answer types. Each problem includes three randomized versions, with additional versions planned for release as the leading open-source LLMs become saturated in UGMathBench. Furthermore, we propose two key metrics: effective accuracy (EAcc), which measures the percentage of correctly solved problems across all three versions, and reasoning gap (Δ), which assesses reasoning robustness by calculating the difference between the average accuracy across all versions and EAcc. Our extensive evaluation of 23 leading LLMs reveals that the highest EAcc achieved is 56.3% by OpenAI-o1-mini, with large Δ values observed across different models. This highlights the need for future research aimed at developing "large reasoning models" with high EAcc and $\Delta = 0$. We anticipate that the release of UGMathBench, along with its detailed evaluation codes, will serve as a valuable resource to advance the development of LLMs in solving mathematical problems.

2103. Can Knowledge Editing Really Correct Hallucinations?

链接：https://iclr.cc/virtual/2025/poster/28744 abstract： Large Language Models (LLMs) suffer from hallucinations, referring to the non-factual information in generated content, despite their superior capacities across tasks. Meanwhile, knowledge editing has been developed as a new popular paradigm to correct erroneous factual knowledge encoded in LLMs with the advantage of avoiding retraining from scratch. However, a common issue of existing evaluation datasets for knowledge editing is that they do not ensure that LLMs actually generate hallucinated answers to the evaluation questions before editing. When LLMs are evaluated on such datasets after being edited by different techniques, it is hard to directly adopt the performance to assess the effectiveness of different knowledge editing methods in correcting hallucinations. Thus, the fundamental question remains insufficiently validated: Can knowledge editing really correct hallucinations in LLMs? We proposed HalluEditBench to holistically benchmark knowledge editing methods in correcting real-world hallucinations. First, we rigorously construct a massive hallucination dataset with 9 domains, 26 topics and more than 6,000 hallucinations. Then, we assess the performance of knowledge editing methods in a holistic way on five dimensions including Efficacy, Generalization, Portability, Locality, and Robustness. Through HalluEditBench, we have provided new insights into the potentials and limitations of different knowledge editing methods in correcting hallucinations, which could inspire future improvements and facilitate progress in the field of knowledge editing.

2104. Improving Uncertainty Estimation through Semantically Diverse Language Generation

链接：https://iclr.cc/virtual/2025/poster/30224 abstract： Large language models (LLMs) can suffer from hallucinations when generating text. These hallucinations impede various applications in society and industry by making LLMs untrustworthy. Current LLMs generate text in an autoregressive fashion by predicting and appending text tokens. When an LLM is uncertain about the semantic meaning of the next tokens to generate, it is likely to start hallucinating. Thus, it has been suggested that predictive uncertainty is one of the main causes of hallucinations. We introduce Semantically Diverse Language Generation (SDLG) to quantify predictive uncertainty in LLMs. SDLG steers the LLM to generate semantically diverse yet likely alternatives for an initially generated text. This approach provides a precise measure of aleatoric semantic uncertainty, detecting whether the initial text is likely to be hallucinated. Experiments on question-answering tasks demonstrate that SDLG consistently outperforms existing methods while being the most computationally efficient, setting a new standard for uncertainty estimation in LLMs.

2105. F-Fidelity: A Robust Framework for Faithfulness Evaluation of Explainable AI

链接：https://iclr.cc/virtual/2025/poster/29329 abstract： Recent research has developed a number of eXplainable AI (XAI) techniques, such as gradient-based approaches, input perturbation-base methods, and black-box explanation methods. While these XAI techniques can extract meaningful insights from deep learning models, how to properly evaluate them remains an open problem. The most widely used approach is to perturb or even remove what the XAI method considers to be the most important features in an input and observe the changes in the output prediction. This approach, although straightforward, suffers the Out-of-Distribution (OOD) problem as the perturbed samples may no longer follow the original data distribution. A recent method RemOve And Retrain (ROAR) solves the OOD issue by retraining the model with perturbed samples guided by explanations. However, using the model retrained based on XAI methods to evaluate these explainers may cause information leakage and thus lead to unfair comparisons. We propose Fine-tuned Fidelity (F-Fidelity), a robust evaluation framework for XAI, which utilizes i) an explanation-agnostic fine-tuning strategy, thus mitigating the information leakage issue, and ii) a random masking operation that ensures that the removal step does not generate an OOD input. We also design controlled experiments with state-of-the-art (SOTA) explainers and their degraded version to verify the correctness of our framework. We conduct experiments on multiple data modalities, such as images, time series, and natural language. The results demonstrate that F-Fidelity significantly improves upon prior evaluation metrics in recovering the ground-truth ranking of the explainers. Furthermore, we show both theoretically and empirically that, given a faithful explainer, F-Fidelity metric can be used to compute the sparsity of influential input components, i.e., to extract the true explanation size.

2106. Auto-GDA: Automatic Domain Adaptation for Efficient Grounding Verification in Retrieval-Augmented Generation

链接：https://iclr.cc/virtual/2025/poster/27875 abstract： While retrieval-augmented generation (RAG) has been shown to enhance factuality of large language model (LLM) outputs, LLMs still suffer from hallucination, generating incorrect or irrelevant information. A common detection strategy involves prompting the LLM again to assess whether its response is grounded in the retrieved evidence, but this approach is costly. Alternatively, lightweight natural language inference (NLI) models for efficient grounding verification can be used at inference time. While existing pre-trained NLI models offer potential solutions, their performance remains subpar compared to larger models on realistic RAG inputs. RAG inputs are more complex than most datasets used for training NLI models and have characteristics specific to the underlying knowledge base, requiring adaptation of the NLI models to a specific target domain. Additionally, the lack of labeled instances in the target domain makes supervised domain adaptation, e.g., through fine-tuning, infeasible. To address these challenges, we introduce Automatic Generative Domain Adaptation (Auto-GDA). Our framework enables unsupervised domain adaptation through synthetic data generation.Unlike previous methods that rely on handcrafted filtering and augmentation strategies, Auto-GDA employs an iterative process to continuously improve the quality of generated samples using weak labels from less efficient teacher models and discrete optimization to select the most promising augmented samples. Experimental results demonstrate the effectiveness of our approach, with models fine-tuned on synthetic data using Auto-GDA often surpassing the performance of the teacher model and reaching the performance level of LLMs at 10% of their computational cost.

2107. Evaluating Large Language Models through Role-Guide and Self-Reflection: A Comparative Study

链接：https://iclr.cc/virtual/2025/poster/30426 abstract： Large Language Models fine-tuned with Reinforcement Learning from Human Feedback (RLHF-LLMs) can over-rely on aligned preferences without truly gaining self-knowledge, leading to hallucination and biases. If an LLM can better access its knowledge and know what it knows, it can avoid making false or unsupported claims. Therefore, it is crucial to evaluate whether LLMs have the ability to know what they know, as it can help to ensure accuracy and faithfulness in real-world applications. Inspired by research in Educational Psychology, surface learners who don’t really know are easily affected by teacher and peer guidance, we treat LLM as a student, incorporate role guidance in prompts to explore whether LLMs really know. Specifically, we propose a novel strategy called Role-Guided and Self-Reflection (RoSe) to fully assess whether LLM “knows it knows”. We introduce multiple combinations of different roles and strong reminder in prompts combined with self-reflection to explore what local information in prompt LLMs rely on and whether LLMs remain unaffected by external guidance with varying roles. Our findings reveal that LLMs are very sensitive to the strong reminder information. Role guidance can help LLMs reduce their reliance on strong reminder. Meanwhile, LLMs tend to trust the role of authority more when guided by different roles. Following these findings, we propose a double-calibrated strategy with verbalized confidence to extract well-calibrated data from closed-source LLM and fine-tune open-source LLMs. Extensive experiments conducted on fine-tuning open-source LLMs demonstrate the effectiveness of double-calibrated strategy in mitigating the reliance of LLMs on local information. For a thorough comparison, we not only employ public JEC-QA and openBookQA datasets, but also construct EG-QA which contains English Grammar multiple-choice question-answering and 14 key knowledge points for assessing self-knowledge and logical reasoning.

2108. Stochastic Semi-Gradient Descent for Learning Mean Field Games with Population-Aware Function Approximation

链接：https://iclr.cc/virtual/2025/poster/28035 abstract： Mean field games (MFGs) model interactions in large-population multi-agent systems through population distributions. Traditional learning methods for MFGs are based on fixed-point iteration (FPI), where policy updates and induced population distributions are computed separately and sequentially. However, FPI-type methods may suffer from inefficiency and instability due to potential oscillations caused by this forward-backward procedure. In this work, we propose a novel perspective that treats the policy and population as a unified parameter controlling the game dynamics. By applying stochastic parameter approximation to this unified parameter, we develop SemiSGD, a simple stochastic gradient descent (SGD)-type method, where an agent updates its policy and population estimates simultaneously and fully asynchronously. Building on this perspective, we further apply linear function approximation (LFA) to the unified parameter, resulting in the first population-aware LFA (PA-LFA) for learning MFGs on continuous state-action spaces. A comprehensive finite-time convergence analysis is provided for SemiSGD with PA-LFA, including its convergence to the equilibrium for linear MFGs—a class of MFGs with a linear structure concerning the population—under the standard contractivity condition, and to a neighborhood of the equilibrium under a more practical condition. We also characterize the approximation error for non-linear MFGs. We validate our theoretical findings with six experiments on three MFGs.

2109. Solving Differential Equations with Constrained Learning

链接：https://iclr.cc/virtual/2025/poster/30951 abstract： (Partial) differential equations (PDEs) are fundamental tools for describing natural phenomena, making their solution crucial in science and engineering. While traditional methods, such as the finite element method, provide reliable solutions, their accuracy is often tied to the use of computationally intensive fine meshes. Moreover, they do not naturally account for measurements or prior solutions, and any change in the problem parameters requires results to be fully recomputed. Neural network-based approaches, such as physics-informed neural networks and neural operators, offer a mesh-free alternative by directly fitting those models to the PDE solution. They can also integrate prior knowledge and tackle entire families of PDEs by simply aggregating additional training losses. Nevertheless, they are highly sensitive to hyperparameters such as collocation points and the weights associated with each loss. This paper addresses these challenges by developing a science-constrained learning (SCL) framework. It demonstrates that finding a (weak) solution of a PDE is equivalent to solving a constrained learning problem with worst-case losses. This explains the limitations of previous methods that minimize the expected value of aggregated losses. SCL also organically integrates structural constraints (e.g., invariances) and (partial) measurements or known solutions. The resulting constrained learning problems can be tackled using a practical algorithm that yields accurate solutions across a variety of PDEs, neural network architectures, and prior knowledge levels without extensive hyperparameter tuning and sometimes even at a lower computational cost.

2110. Efficient Diffusion Transformer Policies with Mixture of Expert Denoisers for Multitask Learning

链接：https://iclr.cc/virtual/2025/poster/28417 abstract： Diffusion Policies have become widely used in Imitation Learning, offering several appealing properties, such as generating multimodal and discontinuous behavior.As models are becoming larger to capture more complex capabilities, their computational demands increase, as shown by recent scaling laws. Therefore, continuing with the current architectures will present a computational roadblock. To address this gap, we propose Mixture-of-Denoising Experts (MoDE) as a novel policy for Imitation Learning.MoDE surpasses current state-of-the-art Transformer-based Diffusion Policies while enabling parameter-efficient scaling through sparse experts and noise-conditioned routing, reducing both active parameters by 40\% and inference costs by 90\% via expert caching.Our architecture combines this efficient scaling with noise-conditioned self-attention mechanism, enabling more effective denoising across different noise levels. MoDE achieves state-of-the-art performance on 134 tasks in four established imitation learning benchmarks (CALVIN and LIBERO). Notably, by pretraining MoDE on diverse robotics data, we achieve 4.01 on CALVIN ABC and 0.95 on LIBERO-90. It surpasses both CNN-based and Transformer Diffusion Policies by an average of 57% across 4 benchmarks, while using 90\% fewer FLOPs and fewer active parameters compared to default Diffusion Transformer architectures. Furthermore, we conduct comprehensive ablations on MoDE's components, providing insights for designing efficient and scalable Transformer architectures for Diffusion Policies. Code and demonstrations are available at https://mbreuss.github.io/MoDE_Diffusion_Policy.

2111. On Generalization Across Environments In Multi-Objective Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28013 abstract： Real-world sequential decision-making tasks often require balancing trade-offs between multiple conflicting objectives, making Multi-Objective Reinforcement Learning (MORL) an increasingly prominent field of research. Despite recent advances, existing MORL literature has narrowly focused on performance within static environments, neglecting the importance of generalizing across diverse settings. Conversely, existing research on generalization in RL has always assumed scalar rewards, overlooking the inherent multi-objectivity of real-world problems. Generalization in the multi-objective context is fundamentally more challenging, as it requires learning a Pareto set of policies addressing varying preferences across multiple objectives. In this paper, we formalize the concept of generalization in MORL and how it can be evaluated. We then contribute a novel benchmark featuring diverse multi-objective domains with parameterized environment configurations to facilitate future studies in this area. Our baseline evaluations of state-of-the-art MORL algorithms on this benchmark reveals limited generalization capabilities, suggesting significant room for improvement. Our empirical findings also expose limitations in the expressivity of scalar rewards, emphasizing the need for multi-objective specifications to achieve effective generalization. We further analyzed the algorithmic complexities within current MORL approaches that could impede the transfer in performance from the single- to multiple-environment settings. This work fills a critical gap and lays the groundwork for future research that brings together two key areas in reinforcement learning: solving multi-objective decision-making problems and generalizing across diverse environments. We make our code available at https://github.com/JaydenTeoh/MORL-Generalization.

2112. TASAR: Transfer-based Attack on Skeletal Action Recognition

链接：https://iclr.cc/virtual/2025/poster/30195 abstract： Skeletal sequence data, as a widely employed representation of human actions, are crucial in Human Activity Recognition (HAR). Recently, adversarial attacks have been proposed in this area, which exposes potential security concerns, and more importantly provides a good tool for model robustness test. Within this research, transfer-based attack is an important tool as it mimics the real-world scenario where an attacker has no knowledge of the target model, but is under-explored in Skeleton-based HAR (S-HAR). Consequently, existing S-HAR attacks exhibit weak adversarial transferability and the reason remains largely unknown. In this paper, we investigate this phenomenon via the characterization of the loss function. We find that one prominent indicator of poor transferability is the low smoothness of the loss function. Led by this observation, we improve the transferability by properly smoothening the loss when computing the adversarial examples. This leads to the first Transfer-based Attack on Skeletal Action Recognition, TASAR. TASAR explores the smoothened model posterior of pre-trained surrogates, which is achieved by a new post-train Dual Bayesian optimization strategy. Furthermore, unlike existing transfer-based methods which overlook the temporal coherence within sequences, TASAR incorporates motion dynamics into the Bayesian attack, effectively disrupting the spatial-temporal coherence of S-HARs. For exhaustive evaluation, we build the first large-scale robust S-HAR benchmark, comprising 7 S-HAR models, 10 attack methods, 3 S-HAR datasets and 2 defense models. Extensive results demonstrate the superiority of TASAR. Our benchmark enables easy comparisons for future studies, with the code available in the https://github.com/yunfengdiao/Skeleton-Robustness-Benchmark.

2113. Diffusion Policy Policy Optimization

链接：https://iclr.cc/virtual/2025/poster/28475 abstract： We introduce Diffusion Policy Policy Optimization, DPPO, an algorithmic framework including best practices for fine-tuning diffusion-based policies (e.g. Diffusion Policy) in continuous control and robot learning tasks using the policy gradient (PG) method from reinforcement learning (RL). PG methods are ubiquitous in training RL policies with other policy parameterizations; nevertheless, they had been conjectured to be less efficient for diffusion-based policies. Surprisingly, we show that DPPO achieves the strongest overall performance and efficiency for fine-tuning in common benchmarks compared to other RL methods for diffusion-based policies and also compared to PG fine-tuning of other policy parameterizations. Through experimental investigation, we find that DPPO takes advantage of unique synergies between RL fine-tuning and the diffusion parameterization, leading to structured and on-manifold exploration, stable training, and strong policy robustness. We further demonstrate the strengths of DPPO in a range of realistic settings, including simulated robotic tasks with pixel observations, and via zero-shot deployment of simulation-trained policies on robot hardware in a long-horizon, multi-stage manipulation task.

2114. Optimizing Posterior Samples for Bayesian Optimization via Rootfinding

链接：https://iclr.cc/virtual/2025/poster/30191 abstract： Bayesian optimization devolves the global optimization of a costly objective function to the global optimization of a sequence of acquisition functions. This inner-loop optimization can be catastrophically difficult if it involves posterior sample paths, especially in higher dimensions. We introduce an efficient global optimization strategy for posterior samples based on global rootfinding. It provides gradient-based optimizers with two sets of judiciously selected starting points, designed to combine exploration and exploitation. The number of starting points can be kept small without sacrificing optimization quality. Remarkably, even with just one point from each set, the global optimum is discovered most of the time. The algorithm scales practically linearly to high dimensions, breaking the curse of dimensionality. For Gaussian process Thompson sampling (GP-TS), we demonstrate remarkable improvement in both inner- and outer-loop optimization, surprisingly outperforming alternatives like EI and GP-UCB in most cases. Our approach also improves the performance of other posterior sample-based acquisition functions, such as variants of entropy search. Furthermore, we propose a sample-average formulation of GP-TS, which has a parameter to explicitly control exploitation and can be computed at the cost of one posterior sample.

2115. DEPfold: RNA Secondary Structure Prediction as Dependency Parsing.

链接：https://iclr.cc/virtual/2025/poster/30440 abstract： RNA secondary structure prediction is critical for understanding RNA functionbut remains challenging due to complex structural elements like pseudoknots andlimited training data. We introduce DEPfold, a novel deep learning approach thatre-frames RNA secondary structure prediction as a dependency parsing problem.DEPfold presents three key innovations: (1) a biologically motivated transformation of RNA structures into labeled dependency trees, (2) a biaffine attentionmechanism for joint prediction of base pairings and their types, and (3) an optimaltree decoding algorithm that enforces valid RNA structural constraints. Unlike traditional energy-based methods, DEPfold learns directly from annotated data andleverages pretrained language models to predict RNA structure. We evaluate DEPfold on both within-family and cross-family RNA datasets, demonstrating significant performance improvements over existing methods. DEPfold shows strongperformance in cross-family generalization when trained on data augmented bytraditional energy-based models, outperforming existing methods on the bpRNAnew dataset. This demonstrates DEPfold’s ability to effectively learn structuralinformation beyond what traditional methods capture. Our approach bridges natural language processing (NLP) with RNA biology, providing a computationallyefficient and adaptable tool for advancing RNA structure prediction and analysis

2116. Open-Source vs Close-Source: The Context Utilization Challenge

链接：https://iclr.cc/virtual/2025/poster/37629 abstract： This blog post aims to evaluate how well the most capable open-source long context large language models (LLMs) utilize context, using the Needle In A Haystack test. We adopt the task of chapter summarization for recently published books to minimize data contamination while ensuring a challenging test. Our results show that open-source models still have room to improve in context utilization compared to close-source models..

2117. Aria-MIDI: A Dataset of Piano MIDI Files for Symbolic Music Modeling

链接：https://iclr.cc/virtual/2025/poster/32079 abstract： We introduce an extensive new dataset of MIDI files, created by transcribing audio recordings of piano performances into their constituent notes. The data pipeline we use is multi-stage, employing a language model to autonomously crawl and score audio recordings from the internet based on their metadata, followed by a stage of pruning and segmentation using an audio classifier. The resulting dataset contains over one million distinct MIDI files, comprising roughly 100,000 hours of transcribed audio. We provide an in-depth analysis of our techniques, offering statistical insights, and investigate the content by extracting metadata tags, which we also provide. Dataset available at https://github.com/loubbrad/aria-midi.

2118. Enhancing Language Model Agents using Diversity of Thoughts

链接：https://iclr.cc/virtual/2025/poster/29196 abstract： A popular approach to building agents using Language Models (LMs) involves iteratively prompting the LM, reflecting on its outputs, and updating the input prompts until the desired task is achieved. However, our analysis reveals two key shortcomings in the existing methods: (i) limited exploration of the decision space due to repetitive reflections, which result in redundant inputs, and (ii) an inability to leverage insights from previously solved tasks. To address these issues, we introduce DoT (Diversity of Thoughts), a novel framework that a) explicitly reduces redundant reflections to enhance decision-space exploration, and b) incorporates a task-agnostic memory component to enable knowledge retrieval from previously solved tasks—unlike current approaches that operate in isolation for each task. Through extensive experiments on a suite of programming benchmarks (HumanEval, MBPP, and LeetCodeHardGym) using a variety of LMs, DoT demonstrates up to a $\textbf{10}$% improvement in Pass@1 while maintaining cost-effectiveness. Furthermore, DoT is modular by design. For instance, when the diverse reflection module of DoT is integrated with existing methods like Tree of Thoughts (ToT), we observe a significant $\textbf{13}$% improvement on Game of 24 (one of the main benchmarks of ToT), highlighting the broad applicability and impact of our contributions across various reasoning tasks.

2119. Variational Bayesian Pseudo-Coreset

链接：https://iclr.cc/virtual/2025/poster/31258 abstract： The success of deep learning requires large datasets and extensive training, which can create significant computational challenges. To address these challenges, pseudo-coresets, small learnable datasets that mimic the entire data, have been proposed. Bayesian Neural Networks, which offer predictive uncertainty and probabilistic interpretation for deep neural networks, also face issues with large-scale datasets due to their high-dimensional parameter space. Prior works on Bayesian Pseudo-Coresets (BPC) attempt to reduce the computational load for computing weight posterior distribution by a small number of pseudo-coresets but suffer from memory inefficiency during BPC training and sub-optimal results. To overcome these limitations, we propose Variational Bayesian Pseudo-Coreset (VBPC), a novel approach that utilizes variational inference to efficiently approximate the posterior distribution, reducing memory usage and computational costs while improving performance across benchmark datasets.

2120. ThinK: Thinner Key Cache by Query-Driven Pruning

链接：https://iclr.cc/virtual/2025/poster/28435 abstract： Large Language Models (LLMs) have revolutionized the field of natural language processing, achieving unprecedented performance across a variety of applications. However, their increased computational and memory demands present significant challenges, especially when handling long sequences.This paper focuses on the long-context scenario, addressing the inefficiencies in KV cache memory consumption during inference. Unlike existing approaches that optimize the memory based on the sequence length, we identify substantial redundancy in the channel dimension of the KV cache, as indicated by an uneven magnitude distribution and a low-rank structure in the attention weights.In response, we propose ThinK, a novel query-dependent KV cache pruning method designed to minimize attention weight loss while selectively pruning the least significant channels. Our approach not only maintains or enhances model accuracy but also achieves a reduction in KV cache memory costs by over 20\% compared with vanilla KV cache eviction and quantization methods. For instance, ThinK integrated with KIVI can achieve $2.8\times$ peak memory reduction while maintaining nearly the same quality, enabling a batch size increase from 4$\times$ (with KIVI alone) to 5$\times$ when using a single GPU. Extensive evaluations on the LLaMA and Mistral models across various long-sequence datasets verified the efficiency of ThinK. Our code has been made available at https://github.com/SalesforceAIResearch/ThinK.

2121. Actions Speak Louder Than Words: Rate-Reward Trade-off in Markov Decision Processes

链接：https://iclr.cc/virtual/2025/poster/29211 abstract： The impact of communication on decision-making systems has been extensively studied under the assumption of dedicated communication channels. We instead consider communicating through actions, where the message is embedded into the actions of an agent which interacts with the environment in a Markov decision process (MDP) framework. We conceptualize the MDP environment as a finite-state channel (FSC), where the actions of the agent serve as the channel input, while the states of the MDP observed by another agent (i.e., receiver) serve as the channel output. Here, we treat the environment as a communication channel over which the agent communicates through its actions, while at the same time, trying to maximize its reward. We first characterize the optimal information theoretic trade-off between the average reward and the rate of reliable communication in the infinite-horizon regime. Then, we propose a novel framework to design a joint control/coding policy, termed Act2Comm, which seamlessly embeds messages into actions. From a communication perspective, Act2Comm functions as a learning-based channel coding scheme for non-differentiable FSCs under input-output constraints. From a control standpoint, Act2Comm learns an MDP policy that incorporates communication capabilities, though at the cost of some control performance. Overall, Act2Comm effectively balances the dual objectives of control and communication in this environment. Experimental results validate Act2Comm's capability to enable reliable communication while maintaining a certain level of control performance.

2122. Efficient Inference for Large Language Model-based Generative Recommendation

链接：https://iclr.cc/virtual/2025/poster/30645 abstract：

2123. Scaling up the Banded Matrix Factorization Mechanism for Large Scale Differentially Private ML

链接：https://iclr.cc/virtual/2025/poster/30906 abstract： Correlated noise mechanisms such as DP Matrix Factorization (DP-MF) have proven to be effective alternatives to DP-SGD in large-epsilon few-epoch training regimes. Significant work has been done to find the best correlated noise strategies, and the current state-of-the-art approach is DP-BandMF , which optimally balances the benefits of privacy amplification and noise correlation. Despite it's utility advantages, severe scalability limitations prevent this mechanism from handling large-scale training scenarios where the number of training iterations may be more than \$10^4\$ and the number of model parameters may exceed \$10^7\$. In this work, we present techniques to scale up DP-BandMF along these two dimensions, significantly extending it's reach and enabling it to effectively handle settings with over \$10^6\$ training iterations and \$10^9\$ model parameters, with no utility degradation at smaller scales.

2124. NetMoE: Accelerating MoE Training through Dynamic Sample Placement

链接：https://iclr.cc/virtual/2025/poster/31182 abstract： Mixture of Experts (MoE) is a widely used technique to expand model sizes for better model quality while maintaining the computation cost constant. In a nutshell, an MoE model consists of multiple experts in each model layer and routes the training tokens to only a fixed number of experts rather than all. In distributed training, as experts are distributed among different GPUs, All-to-All communication is necessary to exchange the training tokens among the GPUs after each time of expert routing. Due to the frequent and voluminous data exchanges, All-to-All communication has become a notable challenge to training efficiency.In this paper, we manage to accelerate All-to-All communication in MoE models from the training sample perspective, which is unexplored so far. In particular, we put forward the observation that tokens in the same training sample have certain levels of locality in expert routing. Motivated by this, we develop NetMoE, which takes such locality into account and dynamically rearranges the placement of training samples to minimize All-to-All communication costs. Specifically, we model the All-to-All communication given the sample placement and formulate an integer programming problem to deduce the optimal placement in polynomial time. Experiments with 32 GPUs show that NetMoE achieves a maximum efficiency improvement of \$1.67 \times\$ compared with current MoE training frameworks.

2125. Adversarial Machine Unlearning

链接：https://iclr.cc/virtual/2025/poster/28088 abstract： This paper focuses on the challenge of machine unlearning, aiming to remove the influence of specific training data on machine learning models. Traditionally, the development of unlearning algorithms runs parallel with that of membership inference attacks (MIA), a type of privacy threat to determine whether a data instance was used for training. However, the two strands are intimately connected: one can view machine unlearning through the lens of MIA success with respect to removed data. Recognizing this connection, we propose a game-theoretic framework that integrates MIAs into the design of unlearning algorithms. Specifically, we model the unlearning problem as a Stackelberg game in which an unlearner strives to unlearn specific training data from a model, while an auditor employs MIAs to detect the traces of the ostensibly removed data. Adopting this adversarial perspective allows the utilization of new attack advancements, facilitating the design of unlearning algorithms. Our framework stands out in two ways. First, it takes an adversarial approach and proactively incorporates the attacks into the design of unlearning algorithms. Secondly, it uses implicit differentiation to obtain the gradients that limit the attacker's success, thus benefiting the process of unlearning. We present empirical results to demonstrate the effectiveness of the proposed approach for machine unlearning.

2126. Storybooth: Training-Free Multi-Subject Consistency for Improved

Visual Storytelling

链接：https://iclr.cc/virtual/2025/poster/30098 abstract： Consistent text-to-image generation depicting the *same* subjects across different images has gained significant recent attention due to its widespread applications in the fields of visual-storytelling and multiple-shot video generation. While remarkable, existing methods often require costly finetuning for each subject and struggle to maintain consistency across multiple characters. In this work, we first analyse the reason for these limitations. Our exploration reveals that the primary-issue stems from *self-attention leakage*, which is exacerbated when trying to ensure consistency across multiple-characters. Motivated by these findings, we next propose a simple yet effective *training and optimization-free approach* for improving multiple-character consistency. In particular, we first leverage multi-modal *chain-of-thought* reasoning in order to *apriori* localize the different subjects across the storyboard frames. The final storyboard images are then generated using a modified diffusion model which includes *1) a bounded cross-attention layer* for ensuring adherence to the initially predicted layout, and *2) a bounded cross-frame self-attention layer* for reducing inter-character attention leakage. Furthermore, we also propose a novel *cross-frame token-merging layer* which allows for improved fine-grain consistency for the storyboard characters. Experimental analysis reveals that proposed approach is not only $\times 30$ faster than prior training-based methods (*eg, textual inversion, dreambooth-lora*) but also surpasses the prior *state-of-the-art*, exhibiting improved multi-character consistency and text-to-image alignment performance.

2127. Concept-ROT: Poisoning Concepts in Large Language Models with Model Editing

链接：https://iclr.cc/virtual/2025/poster/29632 abstract： Model editing methods modify specific behaviors of Large Language Models by altering a small, targeted set of network weights and require very little data and compute. These methods can be used for malicious applications such as inserting misinformation or simple trojans that result in adversary-specified behaviors when a trigger word is present. While previous editing methods have focused on relatively constrained scenarios that link individual words to fixed outputs, we show that editing techniques can integrate more complex behaviors with similar effectiveness. We develop Concept-ROT, a model editing-based method that efficiently inserts trojans which not only exhibit complex output behaviors, but also trigger on high-level concepts -- presenting an entirely new class of trojan attacks. Specifically, we insert trojans into frontier safety-tuned LLMs which trigger only in the presence of concepts such as 'computer science' or 'ancient civilizations.' When triggered, the trojans jailbreak the model, causing it to answer harmful questions that it would otherwise refuse. Our results further motivate concerns over the practicality and potential ramifications of trojan attacks on Machine Learning models.

2128. Morphing Tokens Draw Strong Masked Image Models

链接：https://iclr.cc/virtual/2025/poster/29006 abstract： Masked image modeling (MIM) has emerged as a promising approach for pre-training Vision Transformers (ViTs). MIMs predict masked tokens token-wise to recover target signals that are tokenized from images or generated by pre-trained models like vision-language models. While using tokenizers or pre-trained models is viable, they often offer spatially inconsistent supervision even for neighboring tokens, hindering models from learning discriminative representations. Our pilot study identifies spatial inconsistency in supervisory signals and suggests that addressing it can improve representation learning. Building upon this insight, we introduce Dynamic Token Morphing (DTM), a novel method that dynamically aggregates tokens while preserving context to generate contextualized targets, thereby likely reducing spatial inconsistency. DTM is compatible with various SSL frameworks; we showcase significantly improved MIM results, barely introducing extra training costs. Our method facilitates MIM training by using more spatially consistent targets, resulting in improved training trends as evidenced by lower losses. Experiments on ImageNet-1K and ADE20K demonstrate DTM's superiority, which surpasses complex state-of-the-art MIM methods. Furthermore, the evaluation of transfer learning on downstream tasks like iNaturalist, along with extensive empirical studies, supports DTM's effectiveness.

2129. BoneMet: An Open Large-Scale Multi-Modal Murine Dataset for Breast Cancer Bone Metastasis Diagnosis and Prognosis

链接：https://iclr.cc/virtual/2025/poster/29268 abstract： Breast cancer bone metastasis (BCBM) affects women’s health globally, calling for the development of effective diagnosis and prognosis solutions. While deep learning has exhibited impressive capacities across various healthcare domains, its applicability in BCBM diseases is consistently hindered by the lack of an open, large-scale, deep learning-ready dataset. As such, we introduce the Bone Metastasis (BoneMet) dataset, the first large-scale, publicly available, high-resolution medical resource, which is derived from a well-accepted murine BCBM model. The unique advantage of BoneMet over existing human datasets is repeated sequential scans per subject over the entire disease development phases. The dataset consists of over 67 terabytes of multi-modal medical data, including 2D X-ray images, 3D CT scans, and detailed biological data (e.g., medical records and bone quantitative analysis), collected from more than five hundreds mice spanning from 2019 to 2024. Our BoneMet dataset is well-organized into six components, i.e., RotationX-Ray, Recon-CT, Seg-CT, Regist-CT, RoI-CT, and MiceMediRec. We further show that BoneMet can be readily adopted to build versatile, large-scale AI models for managing BCBM diseases in terms of diagnosis using 2D or 3D images, prognosis of bone deterioration, and sparse-angle 3D reconstruction for safe long-term disease monitoring. Our preliminary results demonstrate that BoneMet has the potentials to jump-start the development and fine-tuning of AI-driven solutions prior to their applications to human patients. To facilitate its easy access and wide dissemination, we have created the BoneMet package, providing three APIs that enable researchers to (i) flexibly process and download the BoneMet data filtered by specific time frames; and (ii) develop and train large-scale AI models for precise BCBM diagnosis and prognosis. The BoneMet dataset is officially available on Hugging Face Datasets at https://huggingface.co/datasets/BoneMet/BoneMet. The BoneMet package is available on the Python Package Index (PyPI) at https://pypi.org/project/BoneMet. Code and tutorials are available at https://github.com/Tiankuo528/BoneMet.

2130. Adaptive Camera Sensor for Vision Models

链接：https://iclr.cc/virtual/2025/poster/30215 abstract： Domain shift remains a persistent challenge in deep-learning-based computer vision, often requiring extensive model modifications or large labeled datasets to address. Inspired by human visual perception, which adjusts input quality through corrective lenses rather than over-training the brain, we propose Lens, a novel camera sensor control method that enhances model performance by capturing high-quality images from the model's perspective, rather than relying on traditional human-centric sensor control. Lens is lightweight and adapts sensor parameters to specific models and scenes in real-time. At its core, Lens utilizes VisiT, a training-free, model-specific quality indicator that evaluates individual unlabeled samples at test time using confidence scores, without additional adaptation costs. To validate Lens, we introduce ImageNet-ES Diverse, a new benchmark dataset capturing natural perturbations from varying sensor and lighting conditions. Extensive experiments on both ImageNet-ES and our new ImageNet-ES Diverse show that Lens significantly improves model accuracy across various baseline schemes for sensor control and model modification, while maintaining low latency in image captures. Lens effectively compensates for large model size differences and integrates synergistically with model improvement techniques. Our code and dataset are available at github.com/Edw2n/Lens.git.

2131. Nonlinear multiregion neural dynamics with parametric impulse response communication channels

链接：https://iclr.cc/virtual/2025/poster/29986 abstract： Cognition arises from the coordinated interaction of brain regions with distinct computational roles. Despite improvements in our ability to extract the dynamics underlying circuit computation from population activity recorded in individual areas, understanding how multiple areas jointly support distributed computation remains a challenge. As part of this effort, we propose a multi-region neural dynamics model composed of two building blocks: i) within-region (potentially driven) nonlinear dynamics and ii) communication channels between regions, parameterized through their impulse response. Together, these choices make it possible to learn nonlinear neural population dynamics and understand the flow of information between regions by drawing from the rich literature of linear systems theory. We develop a state noise inversion free variational filtering and learning algorithm for our model and show, through neuroscientifically inspired numerical experiments, how the proposed model can reveal interpretable characterizations of the local computations within and the flow of information between neural populations. We further validate the efficacy of our approach using simultaneous population recordings from areas V1 and V2.

2132. DELIFT: Data Efficient Language model Instruction Fine-Tuning

链接：https://iclr.cc/virtual/2025/poster/30319 abstract： Fine-tuning large language models (LLMs) is crucial for task specialization but often becomes resource-intensive due to redundant or uninformative data. Existing data selection methods typically rely either on computationally expensive gradient-based metrics or static embeddings that fail to adapt dynamically to the model's evolving state, thus limiting their practical effectiveness. To address this,we propose DELIFT (Data Efficient Language model Instruction Fine-Tuning), leveraging a novel, computationally efficient utility metric inspired by In-Context Learning (ICL). Our ICL-based metric measures the informational value of each data sample by quantifying its effectiveness as an in-context example in improving model predictions for other samples, reflecting its actual contribution relative to the model's current state. Integrated with tailored submodular optimization methods, DELIFT systematically selects diverse, informative subsets optimized specifically for each fine-tuning stage: instruction tuning, task-specific adaptation, and continual fine-tuning. Experimental results across multiple datasets and model scales show DELIFT reduces fine-tuning data requirements by up to 70% without compromising performance, consistently outperforming existing methods by up to 26% in effectiveness and efficiency.

2133. On the Almost Sure Convergence of the Stochastic Three Points Algorithm

链接：https://iclr.cc/virtual/2025/poster/29896 abstract：

2134. On-the-fly Preference Alignment via Principle-Guided Decoding

链接：https://iclr.cc/virtual/2025/poster/29035 abstract：

2135. Normed Spaces for Graph Embedding

链接：https://iclr.cc/virtual/2025/poster/31496 abstract：

2136. Asymptotic Analysis of Two-Layer Neural Networks after One Gradient Step under Gaussian Mixtures Data with Structure

链接：https://iclr.cc/virtual/2025/poster/28057 abstract： In this work, we study the training and generalization performance of two-layer neural networks (NNs) after one gradient descent step under structured data modeled by Gaussian mixtures. While previous research has extensively analyzed this model under isotropic data assumption, such simplifications overlook the complexities inherent in real-world datasets. Our work addresses this limitation by analyzing two-layer NNs under Gaussian mixture data assumption in the asymptotically proportional limit, where the input dimension, number of hidden neurons, and sample size grow with finite ratios. We characterize the training and generalization errors by leveraging recent advancements in Gaussian universality. Specifically, we prove that a high-order polynomial model performs equivalent to the non-linear neural networks under certain conditions. The degree of the equivalent model is intricately linked to both the "data spread" and the learning rate employed during one gradient step. Through extensive simulations, we demonstrate the equivalence between the original model and its polynomial counterpart across various regression and classification tasks. Additionally, we explore how different properties of Gaussian mixtures affect learning outcomes. Finally, we illustrate experimental results on Fashion-MNIST classification, indicating that our findings can translate to realistic data.

2137. Dataset Ownership Verification in Contrastive Pre-trained Models

链接：https://iclr.cc/virtual/2025/poster/27659 abstract： High-quality open-source datasets, which necessitate substantial efforts for curation, has become the primary catalyst for the swift progress of deep learning. Concurrently, protecting these datasets is paramount for the well-being of the data owner. Dataset ownership verification emerges as a crucial method in this domain, but existing approaches are often limited to supervised models and cannot be directly extended to increasingly popular unsupervised pre-trained models. In this work, we propose the first dataset ownership verification method tailored specifically for self-supervised pre-trained models by contrastive learning. Its primary objective is to ascertain whether a suspicious black-box backbone has been pre-trained on a specific unlabeled dataset, aiding dataset owners in upholding their rights. The proposed approach is motivated by our empirical insights that when models are trained with the target dataset, the unary and binary instance relationships within the embedding space exhibit significant variations compared to models trained without the target dataset. We validate the efficacy of this approach across multiple contrastive pre-trained models including SimCLR, BYOL, SimSiam, MOCO v3, and DINO. The results demonstrate that our method rejects the null hypothesis with a $p$-value markedly below $0.05$, surpassing all previous methodologies. Our code is available at https://github.com/xieyc99/DOV4CL.

2138. Equivariant Symmetry Breaking Sets

链接：https://iclr.cc/virtual/2025/poster/31458 abstract： Equivariant neural networks (ENNs) have been shown to be extremely effective in applications involving underlying symmetries. By construction ENNs cannot produce lower symmetry outputs given a higher symmetry input. However, symmetry breaking occurs in many physical systems and we may obtain a less symmetric stable state from an initial highly symmetric one. Hence, it is imperative that we understand how to systematically break symmetry in ENNs. In this work, we propose a novel symmetry breaking framework that is fully equivariant and is the first which fully addresses spontaneous symmetry breaking. We emphasize that our approach is general and applicable to equivariance under any group. To achieve this, we introduce the idea of symmetry breaking sets (SBS). Rather than redesign existing networks, we design sets of symmetry breaking objects which we feed into our network based on the symmetry of our inputs and outputs. We show there is a natural way to define equivariance on these sets, which gives an additional constraint. Minimizing the size of these sets equates to data efficiency. We prove that minimizing these sets translates to a well studied group theory problem, and tabulate solutions to this problem for the point groups. Finally, we provide some examples of symmetry breaking to demonstrate how our approach works in practice. The code for these examples is available at \url{https://github.com/atomicarchitects/equivariant-SBS}.

2139. Size-Generalizable RNA Structure Evaluation by Exploring Hierarchical Geometries

链接：https://iclr.cc/virtual/2025/poster/29684 abstract： Understanding the 3D structure of RNA is essential for deciphering its function and developing RNA-based therapeutics. Geometric Graph Neural Networks (GeoGNNs) that conform to the $\mathrm{E}(3)$-symmetry have advanced RNA structure evaluation, a crucial step toward RNA structure prediction. However, existing GeoGNNs are still defective in two aspects: 1. inefficient or incapable of capturing the full geometries of RNA; 2. limited generalization ability when the size of RNA significantly differs between training and test datasets. In this paper, we propose EquiRNA, a novel equivariant GNN model by exploring the three-level hierarchical geometries of RNA. At its core, EquiRNA effectively addresses the size generalization challenge by reusing the representation of nucleotide, the common building block shared across RNAs of varying sizes. Moreover, by adopting a scalarization-based equivariant GNN as the backbone, our model maintains directional information while offering higher computational efficiency compared to existing GeoGNNs. Additionally, we propose a size-insensitive $K$-nearest neighbor sampling strategy to enhance the model's robustness to RNA size shifts. We test our approach on our created benchmark as well as an existing dataset. The results show that our method significantly outperforms other state-of-the-art methods, providing a robust baseline for RNA 3D structure modeling and evaluation.

2140. Semialgebraic Neural Networks: From roots to representations

链接：https://iclr.cc/virtual/2025/poster/27662 abstract： Many numerical algorithms in scientific computing—particularly in areas like numerical linear algebra, PDE simulation, and inverse problems—produce outputs that can be represented by semialgebraic functions; that is, the graph of the computed function can be described by finitely many polynomial equalities and inequalities. In this work, we introduce Semialgebraic Neural Networks (SANNs), a neural network architecture capable of representing any bounded semialgebraic function, and computing such functions up to the accuracy of a numerical ODE solver chosen by the programmer. Conceptually, we encode the graph of the learned function as the kernel of a piecewise polynomial selected from a class of functions whose roots can be evaluated using a particular homotopy continuation method. We show by construction that the SANN architecture is able to execute this continuation method, thus evaluating the learned semialgebraic function. Furthermore, the architecture can exactly represent even discontinuous semialgebraic functions by executing a continuation method on each connected component of the target function. Lastly, we provide example applications of these networks and show they can be trained with traditional deep-learning techniques.

2141. Geometry-aware RL for Manipulation of Varying Shapes and Deformable Objects

链接：https://iclr.cc/virtual/2025/poster/30840 abstract： Manipulating objects with varying geometries and deformable objects is a major challenge in robotics. Tasks such as insertion with different objects or cloth hanging require precise control and effective modelling of complex dynamics. In this work, we frame this problem through the lens of a heterogeneous graph that comprises smaller sub-graphs, such as actuators and objects, accompanied by different edge types describing their interactions. This graph representation serves as a unified structure for both rigid and deformable objects tasks, and can be extended further to tasks comprising multiple actuators. To evaluate this setup, we present a novel and challenging reinforcement learning benchmark, including rigid insertion of diverse objects, as well as rope and cloth manipulation with multiple end-effectors. These tasks present a large search space, as both the initial and target configurations are uniformly sampled in 3D space. To address this issue, we propose a novel graph-based policy model, dubbed Heterogeneous Equivariant Policy (HEPi), utilizing $SE(3)$ equivariant message passing networks as the main backbone to exploit the geometric symmetry. In addition, by modeling explicit heterogeneity, HEPi can outperform Transformer-based and non-heterogeneous equivariant policies in terms of average returns, sample efficiency, and generalization to unseen objects. Our project page is available at https://thobotics.github.io/hepi.

2142. Predictive Inverse Dynamics Models are Scalable Learners for Robotic Manipulation

链接：https://iclr.cc/virtual/2025/poster/28455 abstract： Current efforts to learn scalable policies in robotic manipulation primarily fall into two categories: one focuses on "action," which involves behavior cloning from extensive collections of robotic data, while the other emphasizes "vision," enhancing model generalization by pre-training representations or generative models, also referred to as world models, using large-scale visual datasets. This paper presents an end-to-end paradigm that predicts actions using inverse dynamics models conditioned on the robot's forecasted visual states, named Predictive Inverse Dynamics Models (PIDM). By closing the loop between vision and action, the end-to-end PIDM can be a better scalable action learner. In practice, we use Transformers to process both visual states and actions, naming the model Seer. It is initially pre-trained on large-scale robotic datasets, such as DROID, and can be adapted to real-world scenarios with a little fine-tuning data. Thanks to large-scale, end-to-end training and the continuous synergy between vision and action at each execution step, Seer significantly outperforms state-of-the-art methods across both simulation and real-world experiments. It achieves improvements of 13% on the LIBERO-LONG benchmark, 22% on CALVIN ABC-D, and 43% in real-world tasks. Notably, it demonstrates superior generalization for novel objects, lighting conditions, and environments under high-intensity disturbances. Code and models will be publicly available.

2143. Nova: Generative Language Models for Assembly Code with Hierarchical Attention and Contrastive Learning

链接：https://iclr.cc/virtual/2025/poster/30979 abstract： Binary code analysis is the foundation of crucial tasks in the security domain; thus building effective binary analysis techniques is more important than ever. Large language models (LLMs) although have brought impressive improvement to source code tasks, do not directly generalize to assembly code due to the unique challenges of assembly: (1) the low information density of assembly and (2) the diverse optimizations in assembly code. To overcome these challenges, this work proposes a hierarchical attention mechanism that builds attention summaries to capture the semantics more effectively and designs contrastive learning objectives to train LLMs to learn assembly optimization. Equipped with these techniques, this work develops Nova, a generative LLM for assembly code. Nova outperforms existing techniques on binary code decompilation by up to 14.84 -- 21.58% higher Pass@1 and Pass@10, and outperforms the latest binary code similarity detection techniques by up to 6.17% Recall@1, showing promising abilities on both assembly generation and understanding tasks.

2144. Towards Synergistic Path-based Explanations for Knowledge Graph

Completion: Exploration and Evaluation

链接：https://iclr.cc/virtual/2025/poster/29360 abstract： Knowledge graph completion (KGC) aims to alleviate the inherent incompleteness of knowledge graphs (KGs), a crucial task for numerous applications such as recommendation systems and drug repurposing. The success of knowledge graph embedding (KGE) models provokes the question about the explainability: ``\textit{Which the patterns of the input KG are most determinant to the prediction}?'' Particularly, path-based explainers prevail in existing methods because of their strong capability for human understanding. In this paper, based on the observation that a fact is usually determined by the synergy of multiple reasoning chains, we propose a novel explainable framework, dubbed KGExplainer, to explore synergistic pathways. KGExplainer is a model-agnostic approach that employs a perturbation-based greedy search algorithm to identify the most crucial synergistic paths as explanations within the local structure of target predictions. To evaluate the quality of these explanations, KGExplainer distills an evaluator from the target KGE model, allowing for the examination of their fidelity. We experimentally demonstrate that the distilled evaluator has comparable predictive performance to the target KGE. Experimental results on benchmark datasets demonstrate the effectiveness of KGExplainer, achieving a human evaluation accuracy of 83.3\% and showing promising improvements in explainability. Code is available at \url{https://github.com/xiaomingaaa/KGExplainer}

2145. Build-A-Scene: Interactive 3D Layout Control for Diffusion-Based Image Generation

链接：https://iclr.cc/virtual/2025/poster/28806 abstract： We propose a diffusion-based approach for Text-to-Image (T2I) generation with interactive 3D layout control.Layout control has been widely studied to alleviate the shortcomings of T2I diffusion models in understanding objects' placement and relationships from text descriptions.Nevertheless, existing approaches for layout control are limited to 2D layouts, require the user to provide a static layout beforehand, and fail to preserve generated images under layout changes.This makes these approaches unsuitable for applications that require 3D object-wise control and iterative refinements, e.g., interior design and complex scene generation. To this end, we leverage the recent advancements in depth-conditioned T2I models and propose a novel approach for interactive 3D layout control.We replace the traditional 2D boxes used in layout control with 3D boxes.Furthermore, we revamp the T2I task as a multi-stage generation process, where at each stage, the user can insert, change, and move an object in 3D while preserving objects from earlier stages.We achieve this through a novel Dynamic Self-Attention (DSA) module and a consistent 3D object translation strategy.To evaluate our approach, we establish a benchmark and an evaluation protocol for interactive 3D layout control.Experiments show that our approach can generate complicated scenes based on 3D layouts, outperforming the standard depth-conditioned T2I methods by two-folds on object generation success rate.Moreover, it outperforms all methods in comparison on preserving objects under layout changes.Project Page: https://abdo-eldesokey.github.io/build-a-scene/

2146. Flow matching achieves almost minimax optimal convergence

链接：https://iclr.cc/virtual/2025/poster/31143 abstract： Flow matching (FM) has gained significant attention as a simulation-free generative model. Unlike diffusion models, which are based on stochastic differential equations, FM employs a simpler approach by solving an ordinary differential equation with an initial condition from a normal distribution, thus streamlining the sample generation process. This paper discusses the convergence properties of FM in terms of the p-Wasserstein distance, a measure of distributional discrepancy. We establish that FM can achieve an almost minimax optimal convergence rate for $1 \leq p \leq 2$, presenting the first theoretical evidence that FM can reach convergence rates comparable to those of diffusion models. Our analysis extends existing frameworks by examining a broader class of mean and variance functions for the vector fields and identifies specific conditions necessary to attain these optimal rates.

2147. LaGeM: A Large Geometry Model for 3D Representation Learning and Diffusion

链接：https://iclr.cc/virtual/2025/poster/30852 abstract： This paper introduces a novel hierarchical autoencoder that maps 3D models into a highly compressed latent space. The hierarchical autoencoder is specifically designed to tackle the challenges arising from large-scale datasets and generative modeling using diffusion. Different from previous approaches that only work on a regular image or volume grid, our hierarchical autoencoder operates on unordered sets of vectors. Each level of the autoencoder controls different geometric levels of detail. We show that the model can be used to represent a wide range of 3D models while faithfully representing high-resolution geometry details. The training of the new architecture takes 0.70x time and 0.58x memory compared to the baseline.We also explore how the new representation can be used for generative modeling. Specifically, we propose a cascaded diffusion framework where each stage is conditioned on the previous stage. Our design extends existing cascaded designs for image and volume grids to vector sets.

2148. Direct Distributional Optimization for Provable Alignment of Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29849 abstract： We introduce a novel alignment method for diffusion models from distribution optimization perspectives while providing rigorous convergence guarantees.We first formulate the problem as a generic regularized loss minimization over probability distributions and directly optimize the distribution using the Dual

Averaging method.Next, we enable sampling from the learned distribution by approximating its score function via Doob's h-transform technique.The proposed framework is supported by rigorous convergence guarantees and an end-to-end bound on the sampling error, which imply that when the original distribution's score is known accurately, the complexity of sampling from shifted distributions is independent of isoperimetric conditions.This framework is broadly applicable to general distribution optimization problems, including alignment tasks in Reinforcement Learning with Human Feedback (RLHF), Direct Preference Optimization (DPO), and Kahneman-Tversky Optimization (KTO). We empirically validate its performance on synthetic and image datasets using the DPO objective.

2149. Enabling Realtime Reinforcement Learning at Scale with Staggered Asynchronous Inference

链接：https://iclr.cc/virtual/2025/poster/28874 abstract： Realtime environments change even as agents perform action inference and learning, thus requiring high interaction frequencies to effectively minimize regret. However, recent advances in machine learning involve larger neural networks with longer inference times, raising questions about their applicability in realtime systems where reaction time is crucial. We present an analysis of lower bounds on regret in realtime reinforcement learning (RL) environments to show that minimizing long-term regret is generally impossible within the typical sequential interaction and learning paradigm, but often becomes possible when sufficient asynchronous compute is available. We propose novel algorithms for staggering asynchronous inference processes to ensure that actions are taken at consistent time intervals, and demonstrate that use of models with high action inference times is only constrained by the environment's effective stochasticity over the inference horizon, and not by action frequency. Our analysis shows that the number of inference processes needed scales linearly with increasing inference times while enabling use of models that are multiple orders of magnitude larger than existing approaches when learning from a realtime simulation of Game Boy games such as Pokemon and Tetris.

2150. Causal Concept Graph Models: Beyond Causal Opacity in Deep Learning

链接：https://iclr.cc/virtual/2025/poster/28500 abstract： Causal opacity denotes the difficulty in understanding the "hidden" causal structure underlying the decisions of deep neural network (DNN) models. This leads to the inability to rely on and verify state-of-the-art DNN-based systems, especially in high-stakes scenarios. For this reason, circumventing causal opacity in DNNs represents a key open challenge at the intersection of deep learning, interpretability, and causality. This work addresses this gap by introducing Causal Concept Graph Models (Causal CGMs), a class of interpretable models whose decision-making process is causally transparent by design. Our experiments show that Causal CGMs can: (i) match the generalisation performance of causally opaque models, (ii) enable human-in-the-loop corrections to mispredicted intermediate reasoning steps, boosting not just downstream accuracy after corrections but also the reliability of the explanations provided for specific instances, and (iii) support the analysis of interventional and counterfactual scenarios, thereby improving the model's causal interpretability and supporting the effective verification of its reliability and fairness.

2151. ADAPT: Attentive Self-Distillation and Dual-Decoder Prediction Fusion for Continual Panoptic Segmentation

链接：https://iclr.cc/virtual/2025/poster/30235 abstract： Panoptic segmentation, which unifies semantic and instance segmentation into a single task, has witnessed considerable success on predefined tasks. However, traditional methods tend to struggle with catastrophic forgetting and poor generalization when learning from a continuous stream of new tasks. While continual learning aims to mitigate these challenges, our study reveals that existing continual panoptic segmentation (CPS) methods often suffer from efficiency or scalability issues. To address these limitations, we propose an efficient adaptation framework that incorporates attentive self-distillation and dual-decoder prediction fusion to efficiently preserve prior knowledge while facilitating model generalization. Specifically, we freeze the majority of model weights, enabling a shared forward pass between the teacher and student models during distillation. Attentive self-distillation then adaptively distills useful knowledge from the old classes without being distracted from non-object regions, which effectively enhances knowledge retention. Additionally, query-level fusion (QLF) is devised to seamlessly integrate the output of the dual decoders without incurring scale inconsistency. Our method achieves state-of-the-art performance on ADE20K and COCO benchmarks. Code is available at https://github.com/Ze-Yang/ADAPT.

2152. Learning to Steer Markovian Agents under Model Uncertainty

链接：https://iclr.cc/virtual/2025/poster/30137 abstract： Designing incentives for an adapting population is a ubiquitous problem in a wide array of economic applications and beyond. In this work, we study how to design additional rewards to steer multi-agent systems towards desired policies \emph{without} prior knowledge of the agents' underlying learning dynamics. Motivated by the limitation of existing works, we consider a new and general category of learning dynamics called \emph{Markovian agents}. We introduce a model-based non-episodic Reinforcement Learning (RL) formulation for our steering problem. Importantly, we focus on learning a \emph{history-dependent} steering strategy to handle the inherent model uncertainty about the agents' learning dynamics. We introduce a novel objective function to encode the desiderata of achieving a good steering outcome with reasonable cost. Theoretically, we identify conditions for the existence of steering strategies to guide agents to the desired policies. Complementing our theoretical contributions, we provide empirical algorithms to approximately solve our objective, which effectively tackles the challenge in learning history-dependent strategies. We demonstrate the efficacy of our algorithms through empirical evaluations.

2153. PvNeXt: Rethinking Network Design and Temporal Motion for Point Cloud Video Recognition

链接：https://iclr.cc/virtual/2025/poster/29195 abstract： Point cloud video perception has become an essential task for the realm of 3D vision. Current 4D representation learning techniques typically engage in iterative processing coupled with dense query operations. Although effective in capturing temporal features, this approach leads to substantial computational redundancy. In this work, we propose a framework, named as PvNeXt, for effective yet efficient point cloud video recognition, via personalized one-shot query operation. Specially, PvNeXt consists of two key modules, the Motion Imitator and the Single-Step Motion Encoder. The former module, the Motion Imitator, is designed to capture the temporal dynamics inherent in sequences of point clouds, thus generating the virtual motion corresponding to each frame. The Single-Step Motion Encoder performs a one-step query operation, associating point cloud of each frame with its corresponding virtual motion frame, thereby extracting motion cues from point cloud sequences and capturing temporal dynamics across the entire sequence. Through the integration of these two modules, {PvNeXt} enables personalized one-shot queries for each frame, effectively eliminating the need for frame-specific looping and intensive query processes. Extensive experiments on multiple benchmarks demonstrate the effectiveness of our method.

2154. Flow With What You Know

链接：https://iclr.cc/virtual/2025/poster/31364 abstract： We provide an accessible introduction to flow-matching and rectified flow models, which are increasingly at the forefront of generative AI applications. Typical descriptions of them are often laden with extensive probability-math equations, which can form barriers to the dissemination and understanding of these models. Fortunately, before they were couched in probabilities, the mechanisms underlying these models were grounded in basic physics, which provides an alternative and highly accessible (yet functionally equivalent) representation of the processes involved.

2155. KLay: Accelerating Arithmetic Circuits for Neurosymbolic AI

链接：https://iclr.cc/virtual/2025/poster/29206 abstract： A popular approach to neurosymbolic AI involves mapping logic formulas to arithmetic circuits (computation graphs consisting of sums and products) and passing the outputs of a neural network through these circuits. This approach enforces symbolic constraints onto a neural network in a principled and end-to-end differentiable way. Unfortunately, arithmetic circuits are challenging to run on modern tensor accelerators as they exhibit a high degree of irregular sparsity. To address this limitation, we introduce knowledge layers (KLay), a new data structure to represent arithmetic circuits that can be efficiently parallelized on GPUs. Moreover, we contribute two algorithms used in the translation of traditional circuit representations to KLay and a further algorithm that exploits parallelization opportunities during circuit evaluations. We empirically show that KLay achieves speedups of multiple orders of magnitude over the state of the art, thereby paving the way towards scaling neurosymbolic AI to larger real-world applications.

2156. Difference-of-submodular Bregman Divergence

链接：https://iclr.cc/virtual/2025/poster/27885 abstract： The Bregman divergence, which is generated from a convex function, is commonly used as a pseudo-distance for comparing vectors or functions in continuous spaces. In contrast, defining an analog of the Bregman divergence for discrete spaces is nontrivial. Iyer & Bilmes (2012b) considered Bregman divergences on discrete domains using submodular functions as generating functions, the discrete analogs of convex functions. In this paper, we further generalize this framework to cases where the generating function is neither submodular nor supermodular, thus increasing the flexibility and representational capacity of the resulting divergence, which we term the difference-of-submodular Bregman divergence. Additionally, we introduce a learnable form of this divergence using permutation-invariant neural networks (NNs) and demonstrate through experiments that it effectively captures key structural properties in discrete data. As a result, the proposed method significantly improves the performance of existing methods on tasks such as clustering and set retrieval problems. This work addresses the challenge of defining meaningful divergences in discrete settings and provides a new tool for tasks requiring structure-preserving distance measures.

2157. Transformers are Universal In-context Learners

链接：https://iclr.cc/virtual/2025/poster/30885 abstract：

2158. Differential learning kinetics govern the transition from memorization to generalization during in-context learning

链接：https://iclr.cc/virtual/2025/poster/30172 abstract：

2159. MMDisCo: Multi-Modal Discriminator-Guided Cooperative Diffusion for Joint Audio and Video Generation

链接：https://iclr.cc/virtual/2025/poster/29154 abstract： This study aims to construct an audio-video generative model with minimal computational cost by leveraging pre-trained single-modal generative models for audio and video.To achieve this, we propose a novel method that guides single-modal models to cooperatively generate well-aligned samples across modalities. Specifically, given two pre-trained base diffusion models, we train a lightweight joint guidance module to adjust scores separately estimated by the base models to match the score of joint distribution over audio and video. We show that this guidance can be computed using the gradient of the optimal discriminator, which distinguishes real audio-video pairs from fake ones independently generated by the base models. Based on this analysis, we construct a joint guidance module by training this discriminator.Additionally, we adopt a loss function to stabilize the discriminator's gradient and make it work as a noise estimator, as in standard diffusion models. Empirical evaluations on several benchmark datasets demonstrate that our method improves both single-modal fidelity and multimodal alignment with relatively few parameters.The code is available at: https://github.com/SonyResearch/MMDisCo.

2160. Improving Convergence Guarantees of Random Subspace Second-order Algorithm for Nonconvex Optimization

链接：https://iclr.cc/virtual/2025/poster/28012 abstract： In recent years, random subspace methods have been actively studied for large-dimensional nonconvex problems. Recent subspace methods have improved theoretical guarantees such as iteration complexity and local convergence rate while reducing computational costs by deriving descent directions in randomly selected low-dimensional subspaces. This paper proposes the Random Subspace Homogenized Trust Region (RSHTR) method with the best theoretical guarantees among random subspace algorithms for nonconvex optimization. RSHTR achieves an ε-approximate first-order stationary point in $O(\varepsilon^{-3/2})$ iterations, converging locally at a linear rate. Furthermore, under rank-deficient conditions, RSHTR satisfies ε-approximate second-order necessary conditions in $O(\varepsilon^{-3/2})$ iterations and exhibits a local quadratic convergence. Experiments on real-world datasets verify the benefits of RSHTR.

2161. GameGen-X: Interactive Open-world Game Video Generation

链接：https://iclr.cc/virtual/2025/poster/30764 abstract： We introduce GameGen-$\mathbb{X}$, the first diffusion transformer model specifically designed for both generating and interactively controlling open-world game videos. This model facilitates high-quality, open-domain generation by approximating various game elements, such as innovative characters, dynamic environments, complex actions, and diverse events. Additionally, it provides interactive controllability, predicting and altering future content based on the current clip, thus allowing for gameplay simulation. To realize this vision, we first collected and built an Open-World Video Game Dataset (OGameData) from scratch. It is the first and largest dataset for open-world game video generation and control, which comprises over one million diverse gameplay video clips with informative captions. GameGen-$\mathbb{X}$ undergoes a two-stage training process, consisting of pre-training and instruction tuning. Firstly, the model was pre-trained via text-to-video generation and video continuation, enabling long-sequence open-domain game video generation with improved fidelity and coherence. Further, to achieve interactive controllability, we designed InstructNet to incorporate game-related multi-modal control signal experts. This allows the model to adjust latent representations based on user inputs, advancing the integration of character interaction and scene content control in video generation. During instruction tuning, only the InstructNet is updated while the pre-trained foundation model is frozen, enabling the integration of interactive controllability without loss of diversity and quality of generated content. GameGen-$\mathbb{X}$ contributes to advancements in open-world game design using generative models. It demonstrates the potential of generative models to serve as auxiliary tools to traditional rendering techniques, demonstrating the potential for merging creative generation with interactive capabilities. The project will be available at https://github.com/GameGen-X/GameGen-X.

2162. Attention in Large Language Models Yields Efficient Zero-Shot Re-Rankers

链接：https://iclr.cc/virtual/2025/poster/27696 abstract： Information retrieval (IR) systems have played a vital role in modern digital life and have cemented their continued usefulness in this new era of generative AI via retrieval-augmented generation. With strong language processing capabilities and remarkable versatility, large language models (LLMs) have become popular choices for zero-shot re-ranking in IR systems. So far, LLM-based re-ranking methods rely on strong generative capabilities, which restricts their use to either specialized or powerful proprietary models. Given these restrictions, we ask: is autoregressive generation necessary and optimal for LLMs to perform re-ranking? We hypothesize that there are abundant signals relevant to re-ranking within LLMs that might not be used to their full potential via generation. To more directly leverage such signals, we propose in-context re-ranking (ICR), a novel method that leverages the change in attention pattern caused by the search query for accurate and efficient re-ranking. We assume that more relevant documents should receive more attention weights when an LLM is processing the query tokens, and leverage such signals for re-ranking. To mitigate the intrinsic biases in LLMs, we propose a calibration method using a content-free query. Due to the absence of generation, ICR only requires two ($O(1)$) forward passes to re-rank N documents, making it substantially more efficient than generative re-ranking methods that require at least $O(N)$ forward passes. Our novel design also enables ICR to be applied to any LLM without specialized training while guaranteeing a well-formed ranking. Extensive experiments with two popular open-weight LLMs on standard single-hop and multi-hop information retrieval benchmarks show that ICR outperforms RankGPT while cutting the latency by more than 60% in practice. Through detailed analyses, we show that ICR's performance is specially strong on tasks that require more complex re-ranking signals, such as handling contextualization and contradiction between the query and passages, as well as information integration across multiple passages. Our findings call for further exploration on novel ways of utilizing open-weight LLMs beyond text generation.

2163. DexTrack: Towards Generalizable Neural Tracking Control for Dexterous Manipulation from Human References

链接：https://iclr.cc/virtual/2025/poster/29153 abstract： We address the challenge of developing a generalizable neural tracking controller for dexterous manipulation from human references. This controller aims to manage a dexterous robot hand to manipulate diverse objects for various purposes defined by kinematic human-object interactions. Developing such a controller is complicated by the intricate contact dynamics of dexterous manipulation and the need for adaptivity, generalizability, and robustness. Current reinforcement learning and trajectory optimization methods often fall short due to their dependence on task-specific rewards or precise system models. We introduce an approach that curates large-scale successful robot tracking demonstrations, comprising pairs of human references and robot actions, to train a neural controller. Utilizing a data flywheel, we iteratively enhance the controller's performance, as well as the number and quality of successful tracking demonstrations. We exploit available tracking demonstrations and carefully integrate reinforcement learning and imitation learning to boost the controller's performance in dynamic environments. At the same time, to obtain high-quality tracking demonstrations, we individually optimize per-trajectory tracking by leveraging the learned tracking controller in a homotopy optimization method. The homotopy optimization, mimicking chain-of-thought, aids in solving challenging trajectory tracking problems to increase demonstration diversity. We showcase our success by training a generalizable neural controller and evaluating it in both simulation and real world. Our method achieves over a 10% improvement in success rates compared to leading baselines. The project website with animated results is available at DexTrack.

2164. ContextGNN: Beyond Two-Tower Recommendation Systems

链接：https://iclr.cc/virtual/2025/poster/28384 abstract： Recommendation systems predominantly utilize two-tower architectures, which evaluate user-item rankings through the inner product of their respective embeddings. However, one key limitation of two-tower models is that they learn a pair-agnostic representation of users and items. In contrast, pair-wise representations either scale poorly due to their quadratic complexity or are too restrictive on the candidate pairs to rank. To address these issues, we introduce Context-based Graph Neural Networks (ContextGNNs), a novel deep learning architecture for link prediction in recommendation systems. The method employs a pair-wise representation technique for familiar items situated within a user's local subgraph, while leveraging two-tower representations to facilitate the recommendation of exploratory items. A final network then predicts how to fuse both pair-wise and two-tower recommendations into a single ranking of items. We demonstrate that ContextGNN is able to adapt to different data characteristics and outperforms existing methods, both traditional and GNN-based, on a diverse set of practical recommendation tasks, improving performance by 20\% on average.

2165. Machine Unlearning Fails to Remove Data Poisoning Attacks

链接：https://iclr.cc/virtual/2025/poster/30218 abstract： We revisit the efficacy of several practical methods for approximate machine unlearning developed for large-scale deep learning. In addition to complying with data deletion requests, one often-cited potential application for unlearning methods is to remove the effects of poisoned data. We experimentally demonstrate that, while existing unlearning methods have been demonstrated to be effective in a number of settings, they fail to remove the effects of data poisoning across a variety of types of poisoning attacks (indiscriminate, targeted, and a newly-introduced Gaussian poisoning attack) and models (image classifiers and LLMs); even when granted a relatively large compute budget. In order to precisely characterize unlearning efficacy, we introduce new evaluation metrics for unlearning based on data poisoning. Our results suggest that a broader perspective, including a wider variety of evaluations, are required to avoid a false sense of confidence in machine unlearning procedures for deep learning without provable guarantees. Moreover, while unlearning methods show some signs of being useful to efficiently remove poisoned data without having to retrain, our work suggests that these methods are not yet ``ready for prime time,'' and currently provide limited benefit over retraining.

2166. SimBa: Simplicity Bias for Scaling Up Parameters in Deep Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28635 abstract： Recent advances in CV and NLP have been largely driven by scaling up the number of network parameters, despite traditional theories suggesting that larger networks are prone to overfitting.These large networks avoid overfitting by integrating components that induce a simplicity bias, guiding models toward simple and generalizable solutions. However, in deep RL, designing and scaling up networks have been less explored.Motivated by this opportunity, we present SimBa, an architecture designed to scale up parameters in deep RL by injecting a simplicity bias. SimBa consists of three components: (i) an observation normalization layer that standardizes inputs with running statistics, (ii) a residual feedforward block to provide a linear pathway from the input to output, and (iii) a layer normalization to control feature magnitudes. By scaling up parameters with SimBa, the sample efficiency of various deep RL algorithms—including off-policy, on-policy, and unsupervised methods—is consistently improved.Moreover, solely by integrating SimBa architecture into SAC, it matches or surpasses state-of-the-art deep RL methods with high computational efficiency across DMC, MyoSuite, and HumanoidBench.These results demonstrate SimBa's broad applicability and effectiveness across diverse RL algorithms and environments.

2167. Vector-ICL: In-context Learning with Continuous Vector Representations

链接：https://iclr.cc/virtual/2025/poster/27769 abstract： Large language models (LLMs) have shown remarkable in-context learning (ICL) capabilities on textual data. We explore whether these capabilities can be extended to continuous vectors from diverse domains, obtained from black-box pretrained encoders. By aligning input data with an LLM's embedding space through lightweight projectors, we observe that LLMs can effectively process and learn from these projected vectors, which we term Vector-ICL. In particular, we find that pretraining projectors with general language modeling objectives enables Vector-ICL, while task-specific finetuning further enhances performance. In our experiments across various tasks and modalities, including text reconstruction, numerical function regression, text classification, summarization, molecule captioning, time-series classification, graph classification, and fMRI decoding, Vector-ICL often surpasses both few-shot ICL and domain-specific model or tuning. We further conduct analyses and case studies, indicating the potential of LLMs to process vector representations beyond traditional token-based paradigms.

2168. TabM: Advancing tabular deep learning with parameter-efficient ensembling

链接：https://iclr.cc/virtual/2025/poster/29590 abstract： Deep learning architectures for supervised learning on tabular data range from simple multilayer perceptrons (MLP) to sophisticated Transformers and retrieval-augmented methods.This study highlights a major, yet so far overlooked opportunity for substantially improving tabular MLPs; namely, parameter-efficient ensembling -- a paradigm for imitating an ensemble of models with just one model.We start by describing TabM -- a simple model based on MLP and BatchEnsemble (an existing technique), improved with our custom modifications.Then, we perform a large scale evaluation of tabular DL architectures on public benchmarks in terms of both task performance and efficiency, which renders the landscape of tabular DL in a new light.In particular, we find that TabM outperforms prior tabular DL models, while the complexity of attention- and retrieval-based methods does not pay off.Lastly, we conduct a detailed empirical analysis, that sheds some light on the high performance of TabM.For example, we show that parameter-efficient ensembling is not an arbitrary trick, but rather a highly effective way to reduce overfitting and improve optimization dynamics of tabular MLPs.Overall, our work brings an impactful technique to tabular DL, analyses its behaviour, and advances the performance-efficiency tradeoff with TabM -- a simple and powerful baseline for researchers and practitioners.

2169. Fourier Sliced-Wasserstein Embedding for Multisets and Measures

链接：https://iclr.cc/virtual/2025/poster/30562 abstract：

2170. Minimax Optimal Reinforcement Learning with Quasi-Optimism

链接：https://iclr.cc/virtual/2025/poster/28715 abstract： In our quest for a reinforcement learning (RL) algorithm that is both practical and provably optimal, we introduce EQO (Exploration via Quasi-Optimism). Unlike existing minimax optimal approaches, EQO avoids reliance on empirical variances and employs a simple bonus term proportional to the inverse of the state-action visit count. Central to EQO is the concept of quasi-optimism, where estimated values need not be fully optimistic, allowing for a simpler yet effective exploration strategy. The algorithm achieves the sharpest known regret bound for tabular RL under the mildest assumptions, proving that fast convergence can be attained with a practical and computationally efficient approach. Empirical evaluations demonstrate that EQO consistently outperforms existing algorithms in both regret performance and computational efficiency, providing the best of both theoretical soundness and practical effectiveness.

2171. EffoVPR: Effective Foundation Model Utilization for Visual Place Recognition

链接：https://iclr.cc/virtual/2025/poster/29874 abstract： The task of Visual Place Recognition (VPR) is to predict the location of a query image from a database of geo-tagged images. Recent studies in VPR have highlighted the significant advantage of employing pre-trained foundation models like DINOv2 for the VPR task. However, these models are often deemed inadequate for VPR without further fine-tuning on VPR-specific data.In this paper, we present an effective approach to harness the potential of a foundation model for VPR. We show that features extracted from self-attention layers can act as a powerful re-ranker for VPR, even in a zero-shot setting. Our method not only outperforms previous zero-shot approaches but also introduces results competitive with several supervised methods.We then show that a single-stage approach utilizing internal ViT layers for pooling can produce global features that achieve state-of-the-art performance, with impressive feature compactness down to 128D. Moreover, integrating our local foundation features for re-ranking further widens this performance gap. Our method also demonstrates exceptional robustness and generalization, setting new state-of-the-art performance, while handling challenging conditions such as occlusion, day-night transitions, and seasonal variations.

2172. RFMamba: Frequency-Aware State Space Model for RF-Based Human-Centric Perception

链接：https://iclr.cc/virtual/2025/poster/28529 abstract： Human-centric perception with radio frequency (RF) signals has recently entered a new era of end-to-end processing with Transformers. Considering the long-sequence nature of RF signals, the State Space Model (SSM) has emerged as a superior alternative due to its effective long-sequence modeling and linear complexity. However, integrating SSM into RF-based sensing presents unique challenges including the fundamentally different signal representation, distinct frequency responses in different scenarios, and incomplete capture caused by specular reflection. To address this, we carefully devise a dual-branch SSM block that is characterized by adaptively grasping the most informative frequency cues and the assistant spatial information to fully explore the human representations from radar echoes. Based on these two branchs, we further introduce an SSM-based network for handling various downstream human perception tasks, named RFMamba. Extensive experimental results demonstrate the superior performance of our proposed RFMamba across all three downstream tasks. To the best of our knowledge, RFMamba is the first attempt to introduce SSM into RF-based human-centric perception.

2173. EC-Diffuser: Multi-Object Manipulation via Entity-Centric Behavior Generation

链接：https://iclr.cc/virtual/2025/poster/28379 abstract： Object manipulation is a common component of everyday tasks, but learning to manipulate objects from high-dimensional observations presents significant challenges. These challenges are heightened in multi-object environments due to the combinatorial complexity of the state space as well as of the desired behaviors. While recent approaches have utilized large-scale offline data to train models from pixel observations, achieving performance gains through scaling, these methods struggle with compositional generalization in unseen object configurations with constrained network and dataset sizes. To address these issues, we propose a novel behavioral cloning (BC) approach that leverages object-centric representations and an entity-centric Transformer with diffusion-based optimization, enabling efficient learning from offline image data. Our method first decomposes observations into Deep Latent Particles (DLP), which are then processed by our entity-centric Transformer that computes attention at the particle level, simultaneously predicting object dynamics and the agent's actions. Combined with the ability of diffusion models to capture multi-modal behavior distributions, this results in substantial performance improvements in multi-object tasks and, more importantly, enables compositional generalization. We present BC agents capable of zero-shot generalization to perform tasks with novel compositions of objects and goals, including larger numbers of objects than seen during training. We provide video rollouts on our webpage: https://sites.google.com/view/ec-diffuser.

2174. Demystifying Topological Message-Passing with Relational Structures: A Case Study on Oversquashing in Simplicial Message-Passing

链接：https://iclr.cc/virtual/2025/poster/29705 abstract： Topological deep learning (TDL) has emerged as a powerful tool for modeling higher-order interactions in relational data. However, phenomena such as oversquashing in topological message-passing remain understudied and lack theoretical analysis. We propose a unifying axiomatic framework that bridges graph and topological message-passing by viewing simplicial and cellular complexes and their message-passing schemes through the lens of relational structures. This approach extends graph-theoretic results and algorithms to higher-order structures, facilitating the analysis and mitigation of oversquashing in topological message-passing networks. Through theoretical analysis and empirical studies on simplicial networks, we demonstrate the potential of this framework to advance TDL.

2175. Revealing and Mitigating Over-Attention in Knowledge Editing

链接：https://iclr.cc/virtual/2025/poster/30995 abstract： Large Language Models~(LLMs) have demonstrated superior performance across a wide range of tasks, but they still exhibit undesirable errors due to incorrect knowledge learned from the training data. To avoid this, knowledge editing methods emerged to precisely edit the specific model knowledge via efficiently modifying a very small percentage of parameters. However, those methods can lead to the problem of Specificity Failure, where the existing knowledge and capabilities are severely degraded due to editing.Our preliminary indicates that Specificity Failure primarily stems from the model's attention heads assigning excessive attention scores to entities related to the edited knowledge, thereby unduly focusing on specific snippets within the context, which we denote as the Attention Drift phenomenon.To mitigate such Attention Drift issue, we introduce a simple yet effective method Selective Attention Drift Restriction(SADR), which introduces an additional regularization term during the knowledge editing process to restrict changes in the attention weight distribution, thereby preventing undue focus on the edited entity.Experiments on five frequently-used strong LLMs demonstrate the effectiveness of our method, where SADR can significantly mitigate Specificity Failure in the predominant knowledge editing tasks.

2176. SCOPE: A Self-supervised Framework for Improving Faithfulness in Conditional Text Generation

链接：https://iclr.cc/virtual/2025/poster/28981 abstract： Large Language Models (LLMs), when used for conditional text generation, often produce hallucinations, i.e., information that is unfaithful or not grounded in the input context. This issue arises in typical conditional text generation tasks, such as text summarization and data-to-text generation, where the goal is to produce fluent text based on contextual input. When fine-tuned on specific domains, LLMs struggle to provide faithful answers to a given context, often adding information or generating errors. One underlying cause of this issue is that LLMs rely on statistical patterns learned from their training data. This reliance can interfere with the model's ability to stay faithful to a provided context, leading to the generation of ungrounded information. We build upon this observation and introduce a novel self-supervised method for generating a training set of unfaithful samples. We then refine the model using a training process that encourages the generation of grounded outputs over unfaithful ones, drawing on preference-based training. Our approach leads to significantly more grounded text generation, outperforming existing self-supervised techniques in faithfulness, as evaluated through automatic metrics, LLM-based assessments, and human evaluations.

2177. SSLAM: Enhancing Self-Supervised Models with Audio Mixtures for Polyphonic Soundscapes

链接：https://iclr.cc/virtual/2025/poster/28347 abstract： Self-supervised pre-trained audio networks have seen widespread adoption in real-world systems, particularly in multi-modal large language models. These networks are often employed in a frozen state, under the assumption that the self-supervised pre-training has sufficiently equipped them to handle real-world audio. However, a critical question remains: how well do these models actually perform in real-world conditions, where audio is typically polyphonic and complex, involving multiple overlapping sound sources? Current audio self-supervised learning (SSL) methods are often benchmarked on datasets predominantly featuring monophonic audio, such as environmental sounds, and speech. As a result, the ability of SSL models to generalize to polyphonic audio, a common characteristic in natural scenarios, remains underexplored. This limitation raises concerns about the practical robustness of SSL models in more realistic audio settings. To address this gap, we introduce Self-Supervised Learning from Audio Mixtures (SSLAM), a novel direction in audio SSL research, designed to improve the model’s ability to learn from polyphonic data while maintaining strong performance on monophonic data. We thoroughly evaluate SSLAM on standard audio SSL benchmark datasets which are predominantly monophonic and conduct a comprehensive comparative analysis against state-of-the-art (SOTA) methods using a range of high-quality, publicly available polyphonic datasets. SSLAM not only improves model performance on polyphonic audio, but also maintains or exceeds performance on standard audio SSL benchmarks. Notably, it achieves up to a 3.9% improvement on the AudioSet-2M(AS-2M), reaching a mean average precision (mAP) of 50.2. For polyphonic datasets, SSLAM sets new SOTA in both linear evaluation and fine-tuning regimes with performance improvements of up to 9.1%(mAP). These results demonstrate SSLAM's effectiveness in both polyphonic and monophonic soundscapes, significantly enhancing the performance of audio SSL models. Code and pre-trained models are available at https://github.com/ta012/SSLAM.

2178. When narrower is better: the narrow width limit of Bayesian parallel branching neural networks

链接：https://iclr.cc/virtual/2025/poster/30498 abstract： The infinite width limit of random neural networks is known to result in Neural Networks as Gaussian Process (NNGP) (Lee et al. (2018)), characterized by task-independent kernels. It is widely accepted that larger network widths contribute to improved generalization (Park et al. (2019)). However, this work challenges this notion by investigating the narrow width limit of the Bayesian Parallel BranchingNeural Network (BPB-NN), an architecture that resembles neural networks with residual blocks. We demonstrate that when the width of a BPB-NN is significantly smaller compared to the number of training examples, each branch exhibits more robust learning due to a symmetry breaking of branches in kernel renormalization. Surprisingly, the performance of a BPB-NN in the narrow width limit is generally superior to or comparable to that achieved in the wide width limit in bias-limited scenarios. Furthermore, the readout norms of each branch in the narrow width limit are mostly independent of the architectural hyperparameters but generally reflective of the nature of the data. We demonstrate such phenomenon primarily in the branching graph neural networks, where each branch represents a different order of convolutions of the graph; we also extend the results to other more general architectures such as the residual-MLP and demonstrate that the narrow width effect is a general feature of the branching networks. Our results characterize a newly defined narrow-width regime for parallel branching networks in general.

2179. MeshMask: Physics-Based Simulations with Masked Graph Neural Networks

链接：https://iclr.cc/virtual/2025/poster/29117 abstract： We introduce a novel masked pre-training technique for graph neural networks (GNNs) applied to computational fluid dynamics (CFD) problems. By randomly masking up to 40\% of input mesh nodes during pre-training, we force the model to learn robust representations of complex fluid dynamics. We pair this masking strategy with an asymmetric encoder-decoder architecture and gated multi-layer perceptrons to further enhance performance. The proposed method achieves state-of-the-art results on seven CFD datasets, including a new challenging dataset of 3D intracranial aneurysm simulations with over 250,000 nodes per mesh. Moreover, it significantly improves model performance and training efficiency across such diverse range of fluid simulation tasks. We demonstrate improvements of up to 60\% in long-term prediction accuracy compared to previous best models, while maintaining similar computational costs. Notably, our approach enables effective pre-training on multiple datasets simultaneously, significantly reducing the time and data required to achieve high performance on new tasks.Through extensive ablation studies, we provide insights into the optimal masking ratio, architectural choices, and training strategies.

2180. Scalable Universal T-Cell Receptor Embeddings from Adaptive Immune Repertoires

链接：https://iclr.cc/virtual/2025/poster/27808 abstract： T cells are a key component of the adaptive immune system, targeting infections, cancers, and allergens with specificity encoded by their T cell receptors (TCRs), and retaining a memory of their targets. High-throughput TCR repertoire sequencing captures a cross-section of TCRs that encode the immune history of any subject, though the data are heterogeneous, high dimensional, sparse, and mostly unlabeled. Sets of TCRs responding to the same antigen, i.e., a protein fragment, co-occur in subjects sharing immune genetics and exposure history. Here, we leverage TCR co-occurrence across a large set of TCR repertoires and employ the GloVe (Pennington et al., 2014) algorithm to derive low-dimensional, dense vector representations (embeddings) of TCRs. We then aggregate these TCR embeddings to generate subject-level embeddings based on observed subject-specific TCR subsets. Further, we leverage random projection theory to improve GloVe's computational efficiency in terms of memory usage and training time. Extensive experimental results show that TCR embeddings targeting the same pathogen have high cosine similarity, and subject-level embeddings encode both immune genetics and pathogenic exposure history.

2181. Improved Algorithms for Kernel Matrix-Vector Multiplication Under Sparsity Assumptions

链接：https://iclr.cc/virtual/2025/poster/27853 abstract： Motivated by the problem of fast processing of attention matrices, we study fast algorithms for computing matrix-vector products for asymmetric Gaussian Kernel matrices $K\in \mathbb{R}^{n\times n}$. K's columns are indexed by a set of n keys $k_1,k_2\ldots, k_n\in \mathbb{R}^d$, rows by a set of n queries $q_1,q_2,\ldots,q_n\in \mathbb{R}^d $, and its i,j entry is $K_{ij} = e^{-\|q_i-k_j\|_2^2/2\sigma^2}$ for some bandwidth parameter $\sigma>0$. Given a vector $x\in \mathbb{R}^n$ and error parameter $\epsilon>0$, our task is to output a $y\in \mathbb{R}^n$ such that $\|Kx-y\|_2\leq \epsilon \|x\|_2$ in time subquadratic in n and linear in d. Our algorithms rely on the following modelling assumption about the matrices K: the sum of the entries of K scales linearly in n, as opposed to worst case quadratic growth. We validate this assumption experimentally, for Gaussian kernel matrices encountered in various settings such as fast attention computation in LLMs. Under this assumption, we obtain the first subquadratic time algorithm for kernel matrix-vector multiplication for unrestricted vectors.

2182. Neural Exploratory Landscape Analysis for Meta-Black-Box-Optimization

链接：https://iclr.cc/virtual/2025/poster/30417 abstract： Recent research in Meta-Black-Box-Optimization (MetaBBO) have shown that meta-trained neural networks can effectively guide the design of black-box optimizers, significantly reducing the need for expert tuning and delivering robust performance across complex problem distributions. Despite their success, a paradox remains: MetaBBO still rely on human-crafted Exploratory Landscape Analysis features to inform the meta-level agent about the low-level optimization progress. To address the gap, this paper proposes Neural Exploratory Landscape Analysis (NeurELA), a novel framework that dynamically profiles landscape features through a two-stage, attention-based neural network, executed in an entirely end-to-end fashion. NeurELA is pre-trained over a variety of MetaBBO algorithms using a multi-task neuroevolution strategy. Extensive experiments show that NeurELA achieves consistently superior performance when integrated into different and even unseen MetaBBO tasks and can be efficiently fine-tuned for further performance boost. This advancement marks a pivotal step in making MetaBBO algorithms more autonomous and broadly applicable. The source code of NeurELA can be accessed at https://anonymous.4open.science/r/Neur-ELA-303C.

2183. NextBestPath: Efficient 3D Mapping of Unseen Environments

链接：https://iclr.cc/virtual/2025/poster/30819 abstract： This work addresses the problem of active 3D mapping, where an agent must find an efficient trajectory to exhaustively reconstruct a new scene.Previous approaches mainly predict the next best view near the agent's location, which is prone to getting stuck in local areas. Additionally, existing indoor datasets are insufficient due to limited geometric complexity and inaccurate ground truth meshes.To overcome these limitations, we introduce a novel dataset AiMDoom with a map generator for the Doom video game, enabling to better benchmark active 3D mapping in diverse indoor environments.Moreover, we propose a new method we call next-best-path (NBP), which predicts long-term goals rather than focusing solely on short-sighted views.The model jointly predicts accumulated surface coverage gains for long-term goals and obstacle maps, allowing it to efficiently plan optimal paths with a unified model.By leveraging online data collection, data augmentation and curriculum learning, NBP significantly outperforms state-of-the-art methods on both the existing MP3D dataset and our AiMDoom dataset, achieving more efficient mapping in indoor environments of varying complexity.

2184. Safety Representations for Safer Policy Learning

链接：https://iclr.cc/virtual/2025/poster/28826 abstract： Reinforcement learning algorithms typically necessitate extensive exploration of the state space to find optimal policies. However, in safety-critical applications, the risks associated with such exploration can lead to catastrophic consequences. Existing safe exploration methods attempt to mitigate this by imposing constraints, which often result in overly conservative behaviours and inefficient learning. Heavy penalties for early constraint violations can trap agents in local optima, deterring exploration of risky yet high-reward regions of the state space. To address this, we introduce a method that explicitly learns state-conditioned safety representations. By augmenting the state features with these safety representations, our approach naturally encourages safer exploration without being excessively cautious, resulting in more efficient and safer policy learning in safety-critical scenarios. Empirical evaluations across diverse environments show that our method significantly improves task performance while reducing constraint violations during training, underscoring its effectiveness in balancing exploration with safety.

2185. Not All Heads Matter: A Head-Level KV Cache Compression Method with Integrated Retrieval and Reasoning

链接：https://iclr.cc/virtual/2025/poster/30348 abstract： Key-Value (KV) caching is a common technique to enhance the computational efficiency of Large Language Models (LLMs), but its memory overhead grows rapidly with input length. Prior work has shown that not all tokens are equally important for text generation, proposing layer-level KV cache compression to selectively retain key information. Recognizing the distinct roles of attention heads in generation, we propose HeadKV, a head-level KV cache compression method, and HeadKV-R2, which leverages a novel contextual reasoning ability estimation for compression. Our approach operates at the level of individual heads, estimating their importance for contextual QA tasks that require both retrieval and reasoning capabilities. Extensive experiments across diverse benchmarks (LongBench, LooGLE), model architectures (e.g., Llama-3-8B-Instruct, Mistral-7B-Instruct), and long-context abilities tests demonstrate that our head-level KV cache compression significantly outperforms strong baselines, particularly in low-resource settings (KV size = 64 & 128). Notably, our method retains just 1.5% of the KV cache while achieving 97% of the performance of the full KV cache on the contextual question answering benchmark.

2186. MotherNet: Fast Training and Inference via Hyper-Network Transformers

链接：https://iclr.cc/virtual/2025/poster/30899 abstract： Foundation models are transforming machine learning across many modalities, with in-context learning replacing classical model training. Recent work on tabular data hints at a similar opportunity to build foundation models for classification for numerical data. However, existing meta-learning approaches can not compete with tree-based methods in terms of inference time. In this paper, we propose MotherNet, a hypernetwork architecture trained on synthetic classification tasks that, once prompted with a never-seen-before training set generates the weights of a trained ``child'' neural-network by in-context learning using a single forward pass. In contrast to most existing hypernetworks that are usually trained for relatively constrained multi-task settings, MotherNet can create models for multiclass classification on arbitrary tabular datasets without any dataset specific gradient descent.The child network generated by MotherNet outperforms neural networks trained using gradient descent on small datasets, and is competitive with predictions by TabPFN and standard ML methods like Gradient Boosting. Unlike a direct application of TabPFN, MotherNet generated networks are highly efficient at inference time.

2187. No Preference Left Behind: Group Distributional Preference Optimization

链接：https://iclr.cc/virtual/2025/poster/29097 abstract： Preferences within a group of people are not uniform but follow a distribution. While existing alignment methods like Direct Preference Optimization (DPO) attempt to steer models to reflect human preferences, they struggle to capture the distributional pluralistic preferences within a group. These methods often skew toward dominant preferences, overlooking the diversity of opinions, especially when conflicting preferences arise. To address this issue, we propose Group Distributional Preference Optimization (GDPO), a novel framework that aligns language models with the distribution of preferences within a group by incorporating the concept of beliefs that shape individual preferences. GDPO calibrates a language model using statistical estimation of the group's belief distribution and aligns the model with belief-conditioned preferences, offering a more inclusive alignment framework than traditional methods. In experiments using both synthetic controllable opinion generation and real-world movie review datasets, we show that DPO fails to align with the targeted belief distributions, while GDPO consistently reduces this alignment gap during training. Additionally, our evaluation metrics demonstrate that GDPO outperforms existing approaches in aligning with group distributional preferences, marking a significant advance in pluralistic alignment.

2188. RankSHAP: Shapley Value Based Feature Attributions for Learning to Rank

链接：https://iclr.cc/virtual/2025/poster/31045 abstract： Numerous works propose post-hoc, model-agnostic explanations for learning to rank, focusing on ordering entities by their relevance to a query through feature attribution methods. However, these attributions often weakly correlate or contradict each other, confusing end users. We adopt an axiomatic game-theoretic approach, popular in the feature attribution community, to identify a set of fundamental axioms that every ranking-based feature attribution method should satisfy. We then introduce Rank-SHAP, extending classical Shapley values to ranking. We evaluate the RankSHAP framework through extensive experiments on two datasets, multiple ranking methods and evaluation metrics. Additionally, a user study confirms RankSHAP’s alignment with human intuition. We also perform an axiomatic analysis of existing rank attribution algorithms to determine their compliance with our proposed axioms. Ultimately, our aim is to equip practitioners with a set of axiomatically backed feature attribution methods for studying IR ranking models, that ensure generality as well as consistency.

2189. Optimistic Games for Combinatorial Bayesian Optimization with Application to Protein Design

链接：https://iclr.cc/virtual/2025/poster/27768 abstract： Bayesian optimization (BO) is a powerful framework to optimize black-box expensive-to-evaluate functions via sequential interactions. In several important problems (e.g. drug discovery, circuit design, neural architecture search, etc.), though, such functions are defined over large $\textit{combinatorial and unstructured}$ spaces. This makes existing BO algorithms not feasible due to the intractable maximization of the acquisition function over these domains. To address this issue, we propose $\textbf{GameOpt}$, a novel game-theoretical approach to combinatorial BO. $\textbf{GameOpt}$ establishes a cooperative game between the different optimization variables, and selects points that are game $\textit{equilibria}$ of an upper confidence bound acquisition function. These are stable configurations from which no variable has an incentive to deviate$-$ analog to local optima in continuous domains. Crucially, this allows us to efficiently break down the complexity of the combinatorial domain into individual decision sets, making $\textbf{GameOpt}$ scalable to large combinatorial spaces. We demonstrate the application of $\textbf{GameOpt}$ to the challenging $\textit{protein design}$ problem and validate its performance on four real-world protein datasets. Each protein can take up to 20^{X} possible configurations, where X is the length of a protein, making standard BO methods infeasible. Instead, our approach iteratively selects informative protein configurations and very quickly discovers highly active protein variants compared to other baselines.

2190. POGEMA: A Benchmark Platform for Cooperative Multi-Agent Pathfinding

链接：https://iclr.cc/virtual/2025/poster/30881 abstract： Multi-agent reinforcement learning (MARL) has recently excelled in solving challenging cooperative and competitive multi-agent problems in various environments, typically involving a small number of agents and full observability. Moreover, a range of crucial robotics-related tasks, such as multi-robot pathfinding, which have traditionally been approached with classical non-learnable methods (e.g., heuristic search), are now being suggested for solution using learning-based or hybrid methods. However, in this domain, it remains difficult, if not impossible, to conduct a fair comparison between classical, learning-based, and hybrid approaches due to the lack of a unified framework that supports both learning and evaluation. To address this, we introduce POGEMA, a comprehensive set of tools that includes a fast environment for learning, a problem instance generator, a collection of predefined problem instances, a visualization toolkit, and a benchmarking tool for automated evaluation. We also introduce and define an evaluation protocol that specifies a range of domain-related metrics, computed based on primary evaluation indicators (such as success rate and path length), enabling a fair multi-fold comparison. The results of this comparison, which involves a variety of state-of-the-art MARL, search-based, and hybrid methods, are presented.

2191. Physiome-ODE: A Benchmark for Irregularly Sampled Multivariate Time-Series Forecasting Based on Biological ODEs

链接：https://iclr.cc/virtual/2025/poster/30863 abstract： State-of-the-art methods for forecasting irregularly sampled time series with missing values predominantly rely on just four datasets and a few small toy examples for evaluation. While ordinary differential equations (ODE) are the prevalent models in science and engineering, a baseline model that forecasts a constant value outperforms ODE-based models from the last five years on three of these existing datasets. This unintuitive finding hampers further research on ODE-based models, a more plausible model family.In this paper, we develop a methodology to generate irregularly sampled multivariate time series (IMTS) datasets from ordinary differentialequations and to select challenging instances via rejection sampling. Using this methodology, we create Physiome-ODE, a large and sophisticated benchmark of IMTS datasets consisting of 50 individual datasets, derived from real-world ordinary differential equations from research in biology. Physiome-ODE is the first benchmark for IMTS forecasting that we are aware of and an order of magnitude larger than the current evaluation setting of four datasets. Using our benchmark Physiome-ODE, we show qualitatively completely different results than those derived from the current four datasets: on Physiome-ODE ODE-based models can play to their strength and our benchmark can differentiate in a meaningful way between different IMTS forecasting models. This way, we expect to give a new impulse to research on ODE-based time series modeling.

2192. Agent-Oriented Planning in Multi-Agent Systems

链接：https://iclr.cc/virtual/2025/poster/30386 abstract： Through the collaboration of multiple LLM-empowered agents possessing diverse expertise and tools, multi-agent systems achieve impressive progress in solving real-world problems. Given the user queries, the meta-agents, serving as the brain within multi-agent systems, are required to decompose the queries into multiple sub-tasks that can be allocated to suitable agents capable of solving them, so-called agent-oriented planning. In this study, we identify three critical design principles of agent-oriented planning, including solvability, completeness, and non-redundancy, to ensure that each sub-task can be effectively resolved, resulting in satisfactory responses to user queries. These principles further inspire us to propose AOP, a novel framework for agent-oriented planning in multi-agent systems, leveraging a fast task decomposition and allocation process followed by an effective and efficient evaluation via a reward model. According to the evaluation results, the meta-agent is also responsible for promptly making necessary adjustments to sub-tasks and scheduling. Besides, we integrate a feedback loop into AOP to further enhance the effectiveness and robustness of such a problem-solving process. Extensive experiments demonstrate the advancement of AOP in solving real-world problems compared to both single-agent systems and existing planning strategies for multi-agent systems. The source code is available at https://github.com/lalaliat/Agent-Oriented-Planning

2193. Differentially private learners for heterogeneous treatment effects

链接：https://iclr.cc/virtual/2025/poster/31170 abstract： Patient data is widely used to estimate heterogeneous treatment effects and understand the effectiveness and safety of drugs. Yet, patient data includes highlysensitive information that must be kept private. In this work, we aim to estimatethe conditional average treatment effect (CATE) from observational data underdifferential privacy. Specifically, we present DP-CATE, a novel framework forCATE estimation that is Neyman-orthogonal and ensures differential privacy of the estimates. Our framework is highly general: it applies to any two-stageCATE meta-learner with a Neyman-orthogonal loss function and any machinelearning model can be used for nuisance estimation. We further provide an extension of our DP-CATE, where we employ RKHS regression to release the completeCATE function while ensuring differential privacy. We demonstrate the effectiveness of DP-CATE across various experiments using synthetic and real-worlddatasets. To the best of our knowledge, we are the first to provide a framework forCATE estimation that is doubly robust and differentially private.

2194. Probabilistic Conformal Prediction with Approximate Conditional Validity

链接：https://iclr.cc/virtual/2025/poster/29865 abstract： We develop a new method for generating prediction sets that combines the flexibility of conformal methods with an estimate of the conditional distribution $\textup{P}_{Y \mid X}$. Existing methods, such as conformalized quantile regression and probabilistic conformal prediction, usually provide only a marginal coverage guarantee. In contrast, our approach extends these frameworks to achieve approximately conditional coverage, which is crucial for many practical applications. Our prediction sets adapt to the behavior of the predictive distribution, making them effective even under high heteroscedasticity. While exact conditional guarantees are infeasible without assumptions on the underlying data distribution, we derive non-asymptotic bounds that depend on the total variation distance of the conditional distribution and its estimate. Using extensive simulations, we show that our method consistently outperforms existing approaches in terms of conditional coverage, leading to more reliable statistical inference in a variety of applications.

2195. STRAP: Robot Sub-Trajectory Retrieval for Augmented Policy Learning

链接：https://iclr.cc/virtual/2025/poster/31008 abstract： Robot learning is witnessing a significant increase in the size, diversity, and complexity of pre-collected datasets, mirroring trends in domains such as natural language processing and computer vision. Many robot learning methods treat such datasets as multi-task expert data and learn a multi-task, generalist policy by training broadly across them. Notably, while these generalist policies can improve the average performance across many tasks, the performance of generalist policies on any one task is often suboptimal due to negative transfer between partitions of the data, compared to task-specific specialist policies. In this work, we argue for the paradigm of training policies during deployment given the scenarios they encounter: rather than deploying pre-trained policies to unseen problems in a zero-shot manner, we non-parametrically retrieve and train models directly on relevant data at test time. Furthermore, we show that many robotics tasks share considerable amounts of low-level behaviors and that retrieval at the "sub"-trajectory granularity enables significantly improved data utilization, generalization, and robustness in adapting policies to novel problems. In contrast, existing full-trajectory retrieval methods tend to underutilize the data and miss out on shared cross-task content. This work proposes STRAP, a technique for leveraging pre-trained vision foundation models and dynamic time warping to retrieve sub-sequences of trajectories from large training corpora in a robust fashion. STRAP outperforms both prior retrieval algorithms and multi-task learning methods in simulated and real experiments, showing the ability to scale to much larger offline datasets in the real world as well as the ability to learn robust control policies with just a handful of real-world demonstrations.

2196. Meta-Dynamical State Space Models for Integrative Neural Data Analysis

链接：https://iclr.cc/virtual/2025/poster/29605 abstract： Learning shared structure across environments facilitates rapid learning and adaptive behavior in neural systems. This has been widely demonstrated and applied in machine learning to train models that are capable of generalizing to novel settings. However, there has been limited work exploiting the shared structure in neural activity during similar tasks for learning latent dynamics from neural recordings.Existing approaches are designed to infer dynamics from a single dataset and cannot be readily adapted to account for statistical heterogeneities across recordings. In this work, we hypothesize that similar tasks admit a corresponding family ofrelated solutions and propose a novel approach for meta-learning this solution space from task-related neural activity of trained animals. Specifically, we capture the variabilities across recordings on a low-dimensional manifold which concisely parametrizes this family of dynamics, thereby facilitating rapid learning of latent dynamics given new recordings. We demonstrate the efficacy of our approach onfew-shot reconstruction and forecasting of synthetic dynamical systems, and neural recordings from the motor cortex during different arm reaching tasks.

2197. Diffusion Transformer Captures Spatial-Temporal Dependencies: A Theory for Gaussian Process Data

链接：https://iclr.cc/virtual/2025/poster/29925 abstract： Diffusion Transformer, the backbone of Sora for video generation, successfully scales the capacity of diffusion models, pioneering new avenues for high-fidelity sequential data generation. Unlike static data such as images, sequential data consists of consecutive data frames indexed by time, exhibiting rich spatial and temporal dependencies. These dependencies represent the underlying dynamic model and are critical to validate the generated data. In this paper, we make the first theoretical step towards bridging diffusion transformers for capturing spatial-temporal dependencies. Specifically, we establish score approximation and distribution estimation guarantees of diffusion transformers for learning Gaussian process data with covariance functions of various decay patterns. We highlight how the spatial-temporal dependencies are captured and affect learning efficiency. Our study proposes a novel transformer approximation theory, where the transformer acts to unroll an algorithm. We support our theoretical results by numerical experiments, providing strong evidence that spatial-temporal dependencies are captured within attention layers, aligning with our approximation theory.

2198. Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling

链接：https://iclr.cc/virtual/2025/poster/29115 abstract： Efficiently modeling sequences with infinite context length has long been a challenging problem. Previous approaches have either suffered from quadratic computational complexity or limited extrapolation ability in length generalization. In thiswork, we present Samba, a simple hybrid architecture that layer-wise combinesMamba, a selective State Space Model (SSM), with Sliding Window Attention(SWA). Samba selectively compresses a given sequence into recurrent hiddenstates while still maintaining the ability to precisely recall recent memories with theattention mechanism. We scale Samba up to 3.8B parameters with 3.2T trainingtokens and demonstrate that it significantly outperforms state-of-the-art modelsacross a variety of benchmarks. Pretrained on sequences of 4K length, Sambashows improved perplexity in context lengths of up to 1M in zero-shot. Whenfinetuned on 4K-length sequences, Samba efficiently extrapolates to a 256K context length with perfect memory recall on the Passkey Retrieval task, and exhibitssuperior retrieval extrapolation on the challenging Phonebook task compared tofull-attention models. As a linear-time sequence model, Samba achieves a 3.73×higher throughput compared to Transformers with grouped-query attention for userprompts of 128K length, and a 3.64× speedup when generating 64K tokens withunlimited streaming.

2199. Provable Convergence Bounds for Hybrid Dynamical Sampling and Optimization

链接：https://iclr.cc/virtual/2025/poster/30347 abstract： Analog dynamical accelerators (DXs) are a growing sub-field in computer architecture research, offering order-of-magnitude gains in power efficiency and latency over traditional digital methods in several machine learning, optimization, and sampling tasks. However, limited-capacity accelerators require hybrid analog/digital algorithms to solve real-world problems, commonly using large-neighborhood local search (LNLS) frameworks. Unlike fully digital algorithms, hybrid LNLS has no non-asymptotic convergence guarantees and no principled hyperparameter selection schemes, particularly limiting cross-device training and inference.In this work, we provide non-asymptotic convergence guarantees for hybrid LNLS by reducing to block Langevin Diffusion (BLD) algorithms.Adapting tools from classical sampling theory, we prove exponential KL-divergence convergence for randomized and cyclic block selection strategies using ideal DXs. With finite device variation, we provide explicit bounds on the 2-Wasserstein bias in terms of step duration, noise strength, and function parameters. Our BLD model provides a key link between established theory and novel computing platforms, and our theoretical results provide a closed-form expression linking device variation, algorithm hyperparameters, and performance.

2200. Overcoming Slow Decision Frequencies in Continuous Control: Model-Based Sequence Reinforcement Learning for Model-Free Control

链接：https://iclr.cc/virtual/2025/poster/27877 abstract： Reinforcement learning (RL) is rapidly reaching and surpassing human-level control capabilities. However, state-of-the-art RL algorithms often require timesteps and reaction times significantly faster than human capabilities, which is impractical in real-world settings and typically necessitates specialized hardware. We introduce Sequence Reinforcement Learning (SRL), an RL algorithm designed to produce a sequence of actions for a given input state, enabling effective control at lower decision frequencies. SRL addresses the challenges of learning action sequences by employing both a model and an actor-critic architecture operating at different temporal scales. We propose a "temporal recall" mechanism, where the critic uses the model to estimate intermediate states between primitive actions, providing a learning signal for each individual action within the sequence. Once training is complete, the actor can generate action sequences independently of the model, achieving model-free control at a slower frequency. We evaluate SRL on a suite of continuous control tasks, demonstrating that it achieves performance comparable to state-of-the-art algorithms while significantly reducing actor sample complexity. To better assess performance across varying decision frequencies, we introduce the Frequency-Averaged Score (FAS) metric. Our results show that SRL significantly outperforms traditional RL algorithms in terms of FAS, making it particularly suitable for applications requiring variable decision frequencies. Furthermore, we compare SRL with model-based online planning, showing that SRL achieves comparable FAS while leveraging the same model during training that online planners use for planning.