3201. PolyPythias: Stability and Outliers across Fifty Language Model Pre-Training Runs

链接：https://iclr.cc/virtual/2025/poster/29092 abstract： The stability of language model pre-training and its effects on downstream performance are still understudied. Prior work shows that the training process can yield significantly different results in response to slight variations in initial conditions, e.g., the random seed.Crucially, the research community still lacks sufficient resources and tools to systematically investigate pre-training stability, particularly for decoder-only language models. We introduce the PolyPythias, a set of 45 new training runs for the Pythia model suite: 9 new seeds across 5 model sizes, from 14M to 410M parameters, resulting in about 7k new checkpoints that we release. Using these new 45 training runs, in addition to the 5 already available, we study the effects of different initial conditions determined by the seed---i.e., parameters' initialisation and data order---on (i) downstream performance, (ii) learned linguistic representations, and (iii) emergence of training phases. In addition to common scaling behaviours, our analyses generally reveal highly consistent training dynamics across both model sizes and initial conditions. Further, the new seeds for each model allow us to identify outlier training runs and delineate their characteristics.Our findings show the potential of using these methods to predict training stability.

3202. Efficient Online Reinforcement Learning Fine-Tuning Need Not Retain Offline Data

链接：https://iclr.cc/virtual/2025/poster/30230 abstract： The modern paradigm in machine learning involves pre-training on diverse data, followed by task-specific fine-tuning. In reinforcement learning (RL), this translates to learning via offline RL on a diverse historical dataset, followed by rapid online RL fine-tuning using interaction data. Most RL fine-tuning methods require continued training on offline data for stability and performance. However, this is undesirable because training on diverse offline data is slow and expensive for large datasets, and should, in principle, also limit the performance improvement possible because of constraints or pessimism on offline data. In this paper, we show that retaining offline data is unnecessary as long as we use a properly-designed online RL approach for fine-tuning offline RL initializations. To build this approach, we start by analyzing the role of retaining offline data in online fine-tuning. We find that continued training on offline data is mostly useful for preventing a sudden divergence in the value function at the onset of fine-tuning, caused by a distribution mismatch between the offline data and online rollouts. This divergence typically results in unlearning and forgetting the benefits of offline pre-training. Our approach, Warm-start RL (WSRL), mitigates the catastrophic forgetting of pre-trained initializations using a very simple idea. WSRL employs a warmup phase that seeds the online RL run with a very small number of rollouts from the pre-trained policy to do fast online RL. The data collected during warmup bridges the distribution mismatch, and helps ``recalibrate'' the offline Q-function to the online distribution, allowing us to completely discard offline data without destabilizing the online RL fine-tuning. We show that WSRL is able to fine-tune without retaining any offline data, and is able to learn faster and attains higher performance than existing algorithms irrespective of whether they do or do not retain offline data.

3203. The Labyrinth of Links: Navigating the Associative Maze of Multi-modal LLMs

链接：https://iclr.cc/virtual/2025/poster/27924 abstract： Multi-modal Large Language Models (MLLMs) have exhibited impressive capability. However, recently many deficiencies of MLLMs have been found compared to human intelligence, $\textit{e.g.}$, hallucination. To drive the MLLMs study, the community dedicated efforts to building larger benchmarks with complex tasks. In this paper, we propose benchmarking an essential but usually overlooked intelligence: $\textbf{association}$, a human's basic capability to link observation and prior practice memory. To comprehensively investigate MLLM's performance on the association, we formulate the association task and devise a standard benchmark based on adjective and verb semantic concepts. Instead of costly data annotation and curation, we propose a convenient $\textbf{annotation-free}$ construction method transforming the general dataset for our association tasks. Simultaneously, we devise a rigorous data refinement process to eliminate confusion in the raw dataset. Building on this database, we establish three levels of association tasks: single-step, synchronous, and asynchronous associations. Moreover, we conduct a comprehensive investigation into the MLLMs' zero-shot association capabilities, addressing multiple dimensions, including three distinct memory strategies, both open-source and closed-source MLLMs, cutting-edge Mixture-of-Experts (MoE) models, and the involvement of human experts. Our systematic investigation shows that current open-source MLLMs consistently exhibit poor capability in our association tasks, even the currently state-of-the-art GPT-4V(vision) also has a significant gap compared to humans. We believe our benchmark would pave the way for future MLLM studies. $\textit{Our data and code are available at:} https://mvig-rhos.com/llm_inception.

3204. Answer, Assemble, Ace: Understanding How LMs Answer Multiple Choice Questions

链接：https://iclr.cc/virtual/2025/poster/30890 abstract： Multiple-choice question answering (MCQA) is a key competence of performant transformer language models that is tested by mainstream benchmarks. However, recent evidence shows that models can have quite a range of performance, particularly when the task format is diversified slightly (such as by shuffling answer choice order). In this work we ask: how do successful models perform formatted MCQA? We employ vocabulary projection and activation patching methods to localize key hidden states that encode relevant information for predicting the correct answer. We find that prediction of a specific answer symbol is causally attributed to a few middle layers, and specifically their multi-head self-attention mechanisms. We show that subsequent layers increase the probability of the predicted answer symbol in vocabulary space, and that this probability increase is associated with a sparse set of attention heads with unique roles. We additionally uncover differences in how different models adjust to alternative symbols. Finally, we demonstrate that a synthetic task can disentangle sources of model error to pinpoint when a model has learned formatted MCQA, and show that logit differences between answer choice tokens continue to grow over the course of training.

3205. Let Me Grok for You: Accelerating Grokking via Embedding Transfer from a Weaker Model

链接：https://iclr.cc/virtual/2025/poster/30991 abstract： "Grokking" is a phenomenon where a neural network first memorizes training data and generalizes poorly, but then suddenly transitions to near-perfect generalization after prolonged training. While intriguing, this delayed generalization phenomenon compromises predictability and efficiency. Ideally, models should generalize directly without delay. To this end, this paper proposes GrokTransfer, a simple and principled method for accelerating grokking in training neural networks, based on the key observation that data embedding plays acrucial role in determining whether generalization is delayed. GrokTransfer first trains a smaller, weaker model to reach a nontrivial (but far from optimal) test performance. Then, the learned input embedding from this weaker model is extracted and used to initialize the embedding in the target, stronger model. We rigorously prove that, on a synthetic XOR task where delayed generalization alwaysoccurs in normal training, GrokTransfer enables the target model to generalize directly without delay. Moreover, we demonstrate that, across empirical studies of different tasks, GrokTransfer effectively reshapes the training dynamics and eliminates delayed generalization, for both fully-connected neural networks and Transformers.

3206. Talking Turns: Benchmarking Audio Foundation Models on Turn-Taking Dynamics

链接：https://iclr.cc/virtual/2025/poster/31129 abstract： The recent wave of audio foundation models (FMs) could provide new capabilities for conversational modeling. However, there have been limited efforts to evaluate these audio FMs comprehensively on their ability to have natural and interactive conversations. To engage in meaningful conversation with the end user, we would want the FMs to additionally perform a fluent succession of turns without too much overlapping speech or long stretches of silence. Inspired by this, we ask whether the recently proposed audio FMs can understand, predict, and perform turn-taking events? To answer this, we propose a novel evaluation protocol that can assess spoken dialog system's turn-taking capabilities using a supervised model as a judge that has been trained to predict turn-taking events in human-human conversations. Using this protocol, we present the first comprehensive user study that evaluates existing spoken dialogue systems on their ability to perform turn-taking events and reveal many interesting insights, such as they sometimes do not understand when to speak up, can interrupt too aggressively and rarely backchannel. We further evaluate multiple open-source and proprietary audio FMs accessible through APIs on carefully curated test benchmarks from Switchboard to measure their ability to understand and predict turn-taking events and identify significant room for improvement. We will open source our evaluation platform to promote the development of advanced conversational AI systems.

3207. TIS-DPO: Token-level Importance Sampling for Direct Preference Optimization With Estimated Weights

链接：https://iclr.cc/virtual/2025/poster/28368 abstract： Direct Preference Optimization (DPO) has been widely adopted for preference alignment of Large Language Models (LLMs) due to its simplicity and effectiveness. However, DPO is derived as a bandit problem in which the whole response is treated as a single arm, ignoring the importance differences between tokens, which may affect optimization efficiency and make it difficult to achieve optimal results.In this work, we propose that the optimal data for DPO has equal expected rewards for each token in winning and losing responses, as there is no difference in token importance. However, since the optimal dataset is unavailable in practice, we propose using the original dataset for importance sampling to achieve unbiased optimization. Accordingly, we propose a token-level importance sampling DPO objective named TIS-DPO that assigns importance weights to each token based on its reward.Inspired by previous works, we estimate the token importance weights using the difference in prediction probabilities from a pair of contrastive LLMs. We explore three methods to construct these contrastive LLMs: (1) guiding the original LLM with contrastive prompts, (2) training two separate LLMs using winning and losing responses, and (3) performing forward and reverse DPO training with winning and losing responses.Experiments show that TIS-DPO significantly outperforms various baseline methods on harmlessness and helpfulness alignment and summarization tasks. We also visualize the estimated weights, demonstrating their ability to identify key token positions.

3208. ParaSolver: A Hierarchical Parallel Integral Solver for Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/31147 abstract： This paper explores the challenge of accelerating the sequential inference process of Diffusion Probabilistic Models (DPMs). We tackle this critical issue from a dynamic systems perspective, in which the inherent sequential nature is transformed into a parallel sampling process. Specifically, we propose a unified framework that generalizes the sequential sampling process of DPMs as solving a system of banded nonlinear equations. Under this generic framework, we reveal that the Jacobian of the banded nonlinear equations system possesses a unit-diagonal structure, enabling further approximation for acceleration. Moreover, we theoretically propose an effective initialization approach for parallel sampling methods. Finally, we construct \textit{ParaSolver}, a hierarchical parallel sampling technique that enhances sampling speed without compromising quality. Extensive experiments show that ParaSolver achieves up to \textbf{12.1× speedup} in terms of wall-clock time. The source code is publicly available at https://github.com/Jianrong-Lu/ParaSolver.git.

3209. NVS-Solver: Video Diffusion Model as Zero-Shot Novel View Synthesizer

链接：https://iclr.cc/virtual/2025/poster/27681 abstract： By harnessing the potent generative capabilities of pre-trained large video diffusion models, we propose a new novel view synthesis paradigm that operates without the need for training. The proposed method adaptively modulates the diffusion sampling process with the given views to enable the creation of visually pleasing results from single or multiple views of static scenes or monocular videos of dynamic scenes. Specifically, built upon our theoretical modeling, we iteratively modulate the score function with the given scene priors represented with warped input views to control the video diffusion process. Moreover, by theoretically exploring the boundary of the estimation error, we achieve the modulation in an adaptive fashion according to the view pose and the number of diffusion steps. Extensive evaluations on both static and dynamic scenes substantiate the significant superiority of our method over state-of-the-art methods both quantitatively and qualitatively. The source code can be found on https://github.com/ZHU-Zhiyu/NVS_Solver.

3210. Identifiable Exchangeable Mechanisms for Causal Structure and Representation Learning

链接：https://iclr.cc/virtual/2025/poster/28605 abstract： Identifying latent representations or causal structures is important for good generalization and downstream task performance. However, both fields developed rather independently.We observe that several structure and representation identifiability methods, particularly those that require multiple environments, rely on exchangeable non--i.i.d. (independent and identically distributed) data.To formalize this connection, we propose the Identifiable Exchangeable Mechanisms (IEM) framework to unify key representation and causal structure learning methods. IEM provides a unified probabilistic graphical model encompassing causal discovery, Independent Component Analysis, and Causal Representation Learning.With the help of the IEM model, we generalize the Causal de Finetti theorem of Guo et al., 2022 by relaxing the necessary conditions for causal structure identification in exchangeable data.We term these conditions cause and mechanism variability, and show how they imply a duality condition in identifiable representation learning, leading to new identifiability results.

3211. Narrowing Information Bottleneck Theory for Multimodal Image-Text Representations Interpretability

链接：https://iclr.cc/virtual/2025/poster/30173 abstract： The task of identifying multimodal image-text representations has garnered increasing attention, particularly with models such as CLIP (Contrastive Language-Image Pretraining), which demonstrate exceptional performance in learning complex associations between images and text. Despite these advancements, ensuring the interpretability of such models is paramount for their safe deployment in real-world applications, such as healthcare. While numerous interpretability methods have been developed for unimodal tasks, these approaches often fail to transfer effectively to multimodal contexts due to inherent differences in the representation structures. Bottleneck methods, well-established in information theory, have been applied to enhance CLIP's interpretability. However, they are often hindered by strong assumptions or intrinsic randomness. To overcome these challenges, we propose the Narrowing Information Bottleneck Theory, a novel framework that fundamentally redefines the traditional bottleneck approach. This theory is specifically designed to satisfy contemporary attribution axioms, providing a more robust and reliable solution for improving the interpretability of multimodal models. In our experiments, compared to state-of-the-art methods, our approach enhances image interpretability by an average of 9\%, text interpretability by an average of 58.83\%, and accelerates processing speed by 63.95\%. Our code is publicly accessible at https://github.com/LMBTough/NIB.

3212. Interaction Asymmetry: A General Principle for Learning Composable Abstractions

链接：https://iclr.cc/virtual/2025/poster/29064 abstract： Learning disentangled representations of concepts and re-composing them in unseen ways is crucial for generalizing to out-of-domain situations. However, the underlying properties of concepts that enable such disentanglement and compositional generalization remain poorly understood. In this work, we propose the principle of interaction asymmetry which states: "Parts of the same concept have more complex interactions than parts of different concepts". We formalize this via block diagonality conditions on the $(n+1)$th order derivatives of the generator mapping concepts to observed data, where different orders of "complexity" correspond to different n. Using this formalism, we prove that interaction asymmetry enables both disentanglement and compositional generalization. Our results unify recent theoretical results for learning concepts of objects, which we show are recovered as special cases with $n=0$ or 1. We provide results for up to $n=2$, thus extending these prior works to more flexible generator functions, and conjecture that the same proof strategies generalize to larger n. Practically, our theory suggests that, to disentangle concepts, an autoencoder should penalize its latent capacity and the interactions between concepts during decoding. We propose an implementation of these criteria using a flexible Transformer-based VAE, with a novel regularizer on the attention weights of the decoder. On synthetic image datasets consisting of objects, we provide evidence that this model can achieve comparable object disentanglement to existing models that use more explicit object-centric priors.

3213. Cross-Entropy Is All You Need To Invert the Data Generating Process

链接：https://iclr.cc/virtual/2025/poster/28737 abstract： Supervised learning has become a cornerstone of modern machine learning, yet a comprehensive theory explaining its effectiveness remains elusive. Empirical phenomena, such as neural analogy-making and the linear representation hypothesis, suggest that supervised models can learn interpretable factors of variation in a linear fashion. Recent advances in self-supervised learning, particularly nonlinear Independent Component Analysis, have shown that these methods can recover latent structures by inverting the data generating process. We extend these identifiability results to parametric instance discrimination, then show how insights transfer to the ubiquitous setting of supervised learning with cross-entropy minimization. We prove that even in standard classification tasks, models learn representations of ground-truth factors of variation up to a linear transformation under a certain DGP. We corroborate our theoretical contribution with a series of empirical studies. First, using simulated data matching our theoretical assumptions, we demonstrate successful disentanglement of latent factors. Second, we show that on DisLib, a widely-used disentanglement benchmark, simple classification tasks recover latent structures up to linear transformations. Finally, we reveal that models trained on ImageNet encode representations that permit linear decoding of proxy factors of variation.Together, our theoretical findings and experiments offer a compelling explanation for recent observations of linear representations, such as superposition in neural networks. This work takes a significant step toward a cohesive theory that accounts for the unreasonable effectiveness of supervised learning.

3214. Let the Code LLM Edit Itself When You Edit the Code

链接：https://iclr.cc/virtual/2025/poster/27645 abstract： In this work, we investigate a typical scenario in code generation where a developer edits existing code in real time and requests a code assistant, e.g., a large language model, to re-predict the next token or next line on the fly. Naively, the LLM needs to re-encode the entire KV cache to provide an accurate prediction. However, this process is computationally expensive, especially when the sequence length is long. Simply encoding the edited subsequence and integrating it to the original KV cache meets the temporal confusion problem, leading to significantly worse performance. We address this efficiency and accuracy trade-off by introducing $\underline{\textbf{P}\text{ositional}\ \textbf{I}\text{ntegrity}\ \textbf{E}\text{ncoding}}$ (PIE). Building upon the rotary positional encoding, PIE first removes the rotary matrices in the Key cache that introduce temporal confusion and then reapplies the correct rotary matrices. This process ensures that positional relationships between tokens are correct and requires only a single round of matrix multiplication. We validate the effectiveness of PIE through extensive experiments on the RepoBench-C-8k dataset, utilizing DeepSeek-Coder models with 1.3B, 6.7B, and 33B parameters. Our evaluation includes three real-world coding tasks: code insertion, code deletion, and multi-place code editing. Results demonstrate that PIE reduces computational overhead by over 85% compared to the standard full recomputation approach across all model sizes and tasks while well approximating the model performance.

3215. Learning to Generate Diverse Pedestrian Movements from Web Videos with Noisy Labels

链接：https://iclr.cc/virtual/2025/poster/32103 abstract： Understanding and modeling pedestrian movements in the real world is crucial for applications like motion forecasting and scene simulation. Many factors influence pedestrian movements, such as scene context, individual characteristics, and goals, which are often ignored by the existing human generation methods. Web videos contain natural pedestrian behavior and rich motion context, but annotating them with pre-trained predictors leads to noisy labels. In this work, we propose learning diverse pedestrian movements from web videos. We first curate a large-scale dataset called CityWalkers that captures diverse real-world pedestrian movements in urban scenes. Then, based on CityWalkers, we propose a generative model called PedGen for diverse pedestrian movement generation. PedGen introduces automatic label filtering to remove the low-quality labels and a mask embedding to train with partial labels. It also contains a novel context encoder that lifts the 2D scene context to 3D and can incorporate various context factors in generating realistic pedestrian movements in urban scenes. Experiments show that PedGen outperforms existing baseline methods for pedestrian movement generation by learning from noisy labels and incorporating the context factors. In addition, PedGen achieves zero-shot generalization in both real-world and simulated environments. The code, model, and data are available at https://genforce.github.io/PedGen/.

3216. In Search of Forgotten Domain Generalization

链接：https://iclr.cc/virtual/2025/poster/30328 abstract： Out-of-Domain (OOD) generalization is the ability of a model trained on one or more domains to generalize to unseen domains. In the ImageNet era of computer vision, evaluation sets for measuring a model's OOD performance were designed to be strictly OOD with respect to style. However, the emergence of foundation models and expansive web-scale datasets has obfuscated this evaluation process, as datasets cover a broad range of domains and risk test domain contamination. In search of the forgotten domain generalization, we create large-scale datasets subsampled from LAION---LAION-Natural and LAION-Rendition---that are strictly OOD to corresponding ImageNet and DomainNet test sets in terms of style. Training CLIP models on these datasets reveals that a significant portion of their performance is explained by in-domain examples. This indicates that the OOD generalization challenges from the ImageNet era still prevail and that training on web-scale data merely creates the illusion of OOD generalization. Furthermore, through a systematic exploration of combining natural and rendition datasets in varying proportions, we identify optimal mixing ratios for model generalization across these domains. Our datasets and results re-enable meaningful assessment of OOD robustness at scale---a crucial prerequisite for improving model robustness.

3217. MetaUrban: An Embodied AI Simulation Platform for Urban Micromobility

链接：https://iclr.cc/virtual/2025/poster/28597 abstract： Public urban spaces such as streetscapes and plazas serve residents and accommodate social life in all its vibrant variations. Recent advances in robotics and embodied AI make public urban spaces no longer exclusive to humans. Food delivery bots and electric wheelchairs have started sharing sidewalks with pedestrians, while robot dogs and humanoids have recently emerged in the street. Micromobility enabled by AI for short-distance travel in public urban spaces plays a crucial component in future transportation systems. It is essential to ensure the generalizability and safety of AI models used for maneuvering mobile machines. In this work, we present MetaUrban, a compositional simulation platform for the AI-driven urban micromobility research. MetaUrban can construct an infinite number of interactive urban scenes from compositional elements, covering a vast array of ground plans, object placements, pedestrians, vulnerable road users, and other mobile agents' appearances and dynamics. We design point navigation and social navigation tasks as the pilot study using MetaUrban for urban micromobility research and establish various baselines of Reinforcement Learning and Imitation Learning. We conduct extensive evaluation across mobile machines, demonstrating that heterogeneous mechanical structures significantly influence the learning and execution of AI policies. We perform a thorough ablation study, showing that the compositional nature of the simulated environments can substantially improve the generalizability and safety of the trained mobile agents. MetaUrban will be made publicly available to provide research opportunities and foster safe and trustworthy embodied AI and micromobility in cities. The code and data have been released.

3218. FIG: Flow with Interpolant Guidance for Linear Inverse Problems

链接：https://iclr.cc/virtual/2025/poster/28851 abstract： Diffusion and flow matching models have recently been used to solve various linear inverse problems in image restoration, such as super-resolution and inpainting. Using a pre-trained diffusion or flow-matching model as a prior, most existing methods modify the reverse-time sampling process by incorporating the likelihood information from the measurement. However, they struggle in challenging scenarios, such as high measurement noise or severe ill-posedness. In this paper, we propose Flow with Interpolant Guidance (FIG), an algorithm where reverse-time sampling is efficiently guided with measurement interpolants through theoretically justified schemes. Experimentally, we demonstrate that FIG efficiently produces highly competitive results on a variety of linear image reconstruction tasks on natural image datasets, especially for challenging tasks. Our code is available at: https://riccizz.github.io/FIG/.

3219. Towards Hierarchical Rectified Flow

链接：https://iclr.cc/virtual/2025/poster/30901 abstract： We formulate a hierarchical rectified flow to model data distributions. It hierarchically couples multiple ordinary differential equations (ODEs) and defines a time-differentiable stochastic process that generates a data distribution from a known source distribution. Each ODE resembles the ODE that is solved in a classic rectified flow, but differs in its domain, i.e., location, velocity, acceleration, etc. Unlike the classic rectified flow formulation, which formulates a single ODE in the location domain and only captures the expected velocity field (sufficient to capture a multi-modal data distribution), the hierarchical rectified flow formulation models the multi-modal random velocity field, acceleration field, etc., in their entirety. This more faithful modeling of the random velocity field enables integration paths to intersect when the underlying ODE is solved during data generation. Intersecting paths in turn lead to integration trajectories that are more straight than those obtained in the classic rectified flow formulation, where integration paths cannot intersect. This leads to modeling of data distributions with fewer neural function evaluations. We empirically verify this on synthetic 1D and 2D data as well as MNIST, CIFAR-10, and ImageNet-32 data. Our code is available at: https://riccizz.github.io/HRF/.

3220. SoundCTM: Unifying Score-based and Consistency Models for Full-band Text-to-Sound Generation

链接：https://iclr.cc/virtual/2025/poster/30028 abstract： Sound content creation, essential for multimedia works such as video games and films, often involves extensive trial-and-error, enabling creators to semantically reflect their artistic ideas and inspirations, which evolve throughout the creation process, into the sound.Recent high-quality diffusion-based Text-to-Sound (T2S) generative models provide valuable tools for creators. However, these models often suffer from slow inference speeds, imposing an undesirable burden that hinders the trial-and-error process.While existing T2S distillation models address this limitation through 1-step generation, the sample quality of 1-step generation remains insufficient for production use.Additionally, while multi-step sampling in those distillation models improves sample quality itself, the semantic content changes due to their lack of deterministic sampling capabilities.Thus, developing a T2S generative model that allows creators to efficiently conduct trial-and-error while producing high-quality sound remains a key challenge.To address these issues, we introduce Sound Consistency Trajectory Models (SoundCTM), which allow flexible transitions between high-quality 1-step sound generation and superior sound quality through multi-step deterministic sampling. This allows creators to efficiently conduct trial-and-error with 1-step generation to semantically align samples with their intention, and subsequently refine sample quality with preserving semantic content through deterministic multi-step sampling.To develop SoundCTM, we reframe the CTM training framework, originally proposed in computer vision, and introduce a novel feature distance using the teacher network for a distillation loss. Additionally, while distilling classifier-free guided trajectories, we introduce a ν-sampling, a new algorithm that offers another source of quality improvement. For the ν-sampling, we simultaneously train both conditional and unconditional student models.For production-level generation, we scale up our model to 1B trainable parameters, making SoundCTM-DiT-1B the first large-scale distillation model in the sound community to achieve both promising high-quality 1-

step and multi-step full-band (44.1kHz) generation.Audio samples are available at \url{https://anonymus-soundctm.github.io/soundctm_iclr/}.

3221. Latent Safety-Constrained Policy Approach for Safe Offline Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29119 abstract： In safe offline reinforcement learning, the objective is to develop a policy that maximizes cumulative rewards while strictly adhering to safety constraints, utilizing only offline data. Traditional methods often face difficulties in balancing these constraints, leading to either diminished performance or increased safety risks. We address these issues with a novel approach that begins by learning a conservatively safe policy through the use of Conditional Variational Autoencoders, which model the latent safety constraints. Subsequently, we frame this as a Constrained Reward-Return Maximization problem, wherein the policy aims to optimize rewards while complying with the inferred latent safety constraints. This is achieved by training an encoder with a reward-Advantage Weighted Regression objective within the latent constraint space. Our methodology is supported by theoretical analysis, including bounds on policy performance and sample complexity. Extensive empirical evaluation on benchmark datasets, including challenging autonomous driving scenarios, demonstrates that our approach not only maintains safety compliance but also excels in cumulative reward optimization, surpassing existing methods. Additional visualizations provide further insights into the effectiveness and underlying mechanisms of our approach.

3222. Large Language Models are Interpretable Learners

链接：https://iclr.cc/virtual/2025/poster/28760 abstract： The trade-off between expressiveness and interpretability remains a core challenge when building human-centric models for classification and decision-making. While symbolic rules offer interpretability, they often lack expressiveness, whereas neural networks excel in performance but are known for being black boxes. This paper shows a combination of Large Language Models (LLMs) and symbolic programs can bridge this gap. In the proposed LLM-based Symbolic Programs (LSPs), the pretrained LLM with natural language prompts provides a massive set of interpretable modules that can transform raw input into natural language concepts. Symbolic programs then integrate these modules into interpretable decision rules. To train LSPs, we develop a divide-and-conquer approach to incrementally build the program from scratch, where the learning process of each step is guided by LLMs. To evaluate the effectiveness of LSPs in extracting interpretable and accurate knowledge from data, we introduce IL-Bench, a collection of diverse tasks, including both synthetic and real-world scenarios across different modalities. Empirical results demonstrate LSP's superior performance compared to traditional neurosymbolic programs and vanilla automatic prompt tuning methods. Moreover, as the knowledge learned by LSP is a combination of natural language descriptions and symbolic rules, it is easily transferable to humans (interpretable), and other LLMs, and generalizes well to out-of-distribution samples. Our code and benchmark will be released for future research.

3223. AutoG: Towards automatic graph construction from tabular data

链接：https://iclr.cc/virtual/2025/poster/28741 abstract： Recent years have witnessed significant advancements in graph machine learning (GML), with its applications spanning numerous domains. However, the focus of GML has predominantly been on developing powerful models, often overlooking a crucial initial step: constructing suitable graphs from common data formats, such as tabular data.This construction process is fundamental to applying graph-based models, yet it remains largely understudied and lacks formalization.Our research aims to address this gap by formalizing the graph construction problem and proposing an effective solution. We identify two critical challenges to achieve this goal: 1. The absence of dedicated datasets to formalize and evaluate the effectiveness of graph construction methods, and 2. Existing automatic construction methods can only be applied to some specific cases, while tedious human engineering is required to generate high-quality graphs. To tackle these challenges, we present a two-fold contribution.First, we introduce a set of datasets to formalize and evaluate graph construction methods. Second, we propose an LLM-based solution, AutoG, automatically generating high-quality graph schemas without human intervention.The experimental results demonstrate that the quality of constructed graphs is critical to downstream task performance, and AutoG can generate high-quality graphs that rival those produced by human experts. Our code can be accessible from https://github.com/amazon-science/Automatic-Table-to-Graph-Generation.

3224. AHA: A Vision-Language-Model for Detecting and Reasoning Over Failures in Robotic Manipulation

链接：https://iclr.cc/virtual/2025/poster/30106 abstract： Robotic manipulation in open-world settings requires not only task execution but also the ability to detect and learn from failures. While recent advances in vision-language models (VLMs) and large language models (LLMs) have improved robots' spatial reasoning and problem-solving abilities, they still struggle with failure recognition, limiting their real-world applicability. We introduce AHA, an open-source VLM designed to detect and reason about failures in robotic manipulation using natural language. By framing failure detection as a free-form reasoning task, AHA identifies failures and provides detailed, adaptable explanations across different robots, tasks, and environments. We fine-tuned AHA using FailGen, a scalable framework that generates the first large-scale dataset of robotic failure trajectories, the AHA dataset. FailGen achieves this by procedurally perturbing successful demonstrations from simulation. Despite being trained solely on the AHA dataset, AHA generalizes effectively to real-world failure datasets, robotic systems, and unseen tasks. It surpasses the second-best model (GPT-4o in-context learning) by 10.3% and exceeds the average performance of six compared models including five state-of-the-art VLMs by 35.3% across multiple metrics and datasets. We integrate AHA into three manipulation frameworks that utilize LLMs/VLMs for reinforcement learning, task and motion planning, and zero-shot trajectory generation. AHA’s failure feedback enhances these policies' performances by refining dense reward functions, optimizing task planning, and improving sub-task verification, boosting task success rates by an average of 21.4% across all three tasks compared to GPT-4 models. Project page: https://aha-vlm.github.io

3225. Demystifying Online Clustering of Bandits: Enhanced Exploration Under Stochastic and Smoothed Adversarial Contexts

链接：https://iclr.cc/virtual/2025/poster/31041 abstract： The contextual multi-armed bandit (MAB) problem is crucial in sequential decision-making. A line of research, known as online clustering of bandits, extends contextual MAB by grouping similar users into clusters, utilizing shared features to improve learning efficiency. However, existing algorithms, which rely on the upper confidence bound (UCB) strategy, struggle to gather adequate statistical information to accurately identify unknown user clusters. As a result, their theoretical analyses require several strong assumptions about the "diversity" of contexts generated by the environment, leading to impractical settings, complicated analyses, and poor practical performance. Removing these assumptions has been a long-standing open problem in the clustering of bandits literature. In this work, we provide two partial solutions. First, we introduce an additional exploration phase to accelerate the identification of clusters. We integrate this general strategy into both graph-based and set-based algorithms and propose two new algorithms, UniCLUB and UniSCLUB. Remarkably, our algorithms require substantially weaker assumptions and simpler theoretical analyses while achieving superior cumulative regret compared to previous studies. Second, inspired by the smoothed analysis framework, we propose a more practical setting that eliminates the requirement for i.i.d. context generation used in previous studies, thus enhancing the performance of existing algorithms for online clustering of bandits. Extensive evaluations on both synthetic and real-world datasets demonstrate that our proposed algorithms outperform existing approaches.

3226. Grounding by Trying: LLMs with Reinforcement Learning-Enhanced Retrieval

链接：https://iclr.cc/virtual/2025/poster/30573 abstract： The hallucinations of large language models (LLMs) are increasingly mitigated by allowing LLMs to search for information and to ground their answers in real sources. Unfortunately, LLMs often struggle with posing the right search queries, especially when dealing with complex or otherwise indirect topics. Observing that LLMs can learn to search for relevant facts by $\textit{trying}$ different queries and learning to up-weight queries that successfully produce relevant results, we introduce $\underline{Le}$arning to $\underline{Re}$trieve by $\underline{T}$rying (LeReT), a reinforcement learning framework that explores search queries and uses preference-based optimization to improve their quality. LeReT can improve the absolute retrieval accuracy by up to 29\% and the downstream generator evaluations by 17\%. The simplicity and flexibility of LeReT allows it to be applied to arbitrary off-the-shelf retrievers and makes it a promising technique for improving general LLM pipelines.

3227. Scalable Discrete Diffusion Samplers: Combinatorial Optimization and Statistical Physics

链接：https://iclr.cc/virtual/2025/poster/28285 abstract： Learning to sample from complex unnormalized distributions over discrete domains emerged as a promising research direction with applications in statistical physics, variational inference, and combinatorial optimization. Recent work has demonstrated the potential of diffusion models in this domain. However, existing methods face limitations in memory scaling and thus the number of attainable diffusion steps since they require backpropagation through the entire generative process. To overcome these limitations we introduce two novel training methods for discrete diffusion samplers, one grounded in the policy gradient theorem and the other one leveraging Self-Normalized Neural Importance Sampling (SN-NIS). These methods yield memory-efficient training and achieve state-of-the-art results in unsupervised combinatorial optimization.Numerous scientific applications additionally require the ability of unbiased sampling. We introduce adaptations of SN-NIS and Neural Markov Chain Monte Carlo that enable for the first time the application of discrete diffusion models to this problem. We validate our methods on Ising model benchmarks and find that they outperform popular autoregressive approaches. Our work opens new avenues for applying diffusion models to a wide range of scientific applications in discrete domains that were hitherto restricted to exact likelihood models.

3228. Perturbation-Restrained Sequential Model Editing

链接：https://iclr.cc/virtual/2025/poster/29099 abstract： Model editing is an emerging field that focuses on updating the knowledge embedded within large language models (LLMs) without extensive retraining. However, current model editing methods significantly compromise the general abilities of LLMs as the number of edits increases, and this trade-off poses a substantial challenge to the continual learning of LLMs. In this paper, we first theoretically analyze that the factor affecting the general abilities in sequential model editing lies in the condition number of the edited matrix. The condition number of a matrix represents its numerical sensitivity, and therefore can be used to indicate the extent to which the original knowledge associations stored in LLMs are perturbed after editing. Subsequently, statistical findings demonstrate that the value of this factor becomes larger as the number of edits increases, thereby exacerbating the deterioration of general abilities. To this end, a framework termed Perturbation Restraint on Upper bouNd for Editing (PRUNE) is proposed, which applies the condition number restraints in sequential editing. These restraints can lower the upper bound on perturbation to edited models, thus preserving the general abilities.Systematically, we conduct experiments employing three editing methods on three LLMs across four downstream tasks.The results show that PRUNE can preserve general abilities while maintaining the editing performance effectively in sequential model editing. The code are available at https://github.com/mjy1111/PRUNE.

3229. Not All Prompts Are Made Equal: Prompt-based Pruning of Text-to-Image Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/31091 abstract： Text-to-image (T2I) diffusion models have demonstrated impressive image generation capabilities. Still, their computational intensity prohibits resource-constrained organizations from deploying T2I models after fine-tuning them on their internal target data. While pruning techniques offer a potential solution to reduce the computational burden of T2I models, static pruning methods use the same pruned model for all input prompts, overlooking the varying capacity requirements of different prompts. Dynamic pruning addresses this issue by utilizing a separate sub-network for each prompt, but it prevents batch parallelism on GPUs. To overcome these limitations, we introduce Adaptive Prompt-Tailored Pruning (APTP), a novel prompt-based pruning method designed for T2I diffusion models. Central to our approach is a prompt router model, which learns to determine the required capacity for an input text prompt and routes it to an architecture code, given a total desired compute budget for prompts. Each architecture code represents a specialized model tailored to the prompts assigned to it, and the number of codes is a hyperparameter. We train the prompt router and architecture codes using contrastive learning, ensuring that similar prompts are mapped to nearby codes. Further, we employ optimal transport to prevent the codes from collapsing into a single one. We demonstrate APTP's effectiveness by pruning Stable Diffusion (SD) V2.1 using CC3M and COCO as target datasets. APTP outperforms the single-model pruning baselines in terms of FID, CLIP, and CMMD scores. Our analysis of the clusters learned by APTP reveals they are semantically meaningful. We also show that APTP can automatically discover previously empirically found challenging prompts for SD, e.g., prompts for generating text images, assigning them to higher capacity codes.

3230. Quality Measures for Dynamic Graph Generative Models

链接：https://iclr.cc/virtual/2025/poster/30758 abstract： Deep generative models have recently achieved significant success in modeling graph data, including dynamic graphs, where topology and features evolve over time. However, unlike in vision and natural language domains, evaluating generative models for dynamic graphs is challenging due to the difficulty of visualizing their output, making quantitative metrics essential. In this work, we develop a new quality metric for evaluating generative models of dynamic graphs. Current metrics for dynamic graphs typically involve discretizing the continuous-evolution of graphs into static snapshots and then applying conventional graph similarity measures. This approach has several limitations: (a) it models temporally related events as i.i.d. samples, failing to capture the non-uniform evolution of dynamic graphs; (b) it lacks a unified measure that is sensitive to both features and topology; (c) it fails to provide a scalar metric, requiring multiple metrics without clear superiority; and (d) it requires explicitly instantiating each static snapshot, leading to impractical runtime demands that hinder evaluation at scale. We propose a novel metric based on the Johnson-Lindenstrauss lemma, applying random projections directly to dynamic graph data. This results in an expressive, scalar, and application-agnostic measure of dynamic graph similarity that overcomes the limitations of traditional methods. We also provide a comprehensive empirical evaluation of metrics for continuous-time dynamic graphs, demonstrating the effectiveness of our approach compared to existing methods. Our implementation is available at https://github.com/ryienh/jl-metric.

3231. Injecting Universal Jailbreak Backdoors into LLMs in Minutes

链接：https://iclr.cc/virtual/2025/poster/29167 abstract： Jailbreak backdoor attacks on LLMs have garnered attention for their effectiveness and stealth. However, existing methods rely on the crafting of poisoned datasets and the time-consuming process of fine-tuning. In this work, we propose JailbreakEdit, a novel jailbreak backdoor injection method that exploits model editing techniques to inject a universal jailbreak backdoor into safety-aligned LLMs with minimal intervention in minutes. JailbreakEdit integrates a multi-node target estimation to estimate the jailbreak space, thus creating shortcuts from the backdoor to this estimated jailbreak space that induce jailbreak actions. Our attack effectively shifts the models' attention by attaching strong semantics to the backdoor, enabling it to bypass internal safety mechanisms. Experimental results show that JailbreakEdit achieves a high jailbreak success rate on jailbreak prompts while preserving generation quality, and safe performance on normal queries. Our findings underscore the effectiveness, stealthiness, and explainability of JailbreakEdit, emphasizing the need for more advanced defense mechanisms in LLMs.

3232. VSTAR: Generative Temporal Nursing for Longer Dynamic Video Synthesis

链接：https://iclr.cc/virtual/2025/poster/30044 abstract： Despite tremendous progress in the field of text-to-video (T2V) synthesis, open-sourced T2V diffusion models struggle to generate longer videos with dynamically varying and evolving content. They tend to synthesize quasi-static videos, ignoring the necessary visual change-over-time implied in the text prompt. Meanwhile, scaling these models to enable longer, more dynamic video synthesis often remains computationally intractable. To tackle this challenge, we introduce the concept of Generative Temporal Nursing (GTN), where we adjust the generative process on the fly during inference to improve control over the temporal dynamics and enable generation of longer videos. We propose a method for GTN, dubbed VSTAR, which consists of two key ingredients: Video Synopsis Prompting (VSP) and Temporal

Attention Regularization (TAR), the latter being our core contribution. Based on a systematic analysis, we discover that the temporal units in pretrained T2V models are crucial to control the video dynamics. Upon this finding, we propose a novel regularization technique to refine the temporal attention, enabling training-free longer video synthesis in a single inference pass. For prompts involving visual progression, we leverage LLMs to generate video synopsis - description of key visual states - based on the original prompt to provide better guidance along the temporal axis. We experimentally showcase the superiority of our method in synthesizing longer, visually appealing videos over open-sourced T2V models.

3233. Reward Guided Latent Consistency Distillation

链接：https://iclr.cc/virtual/2025/poster/31456 abstract： Latent Consistency Distillation (LCD) has emerged as a promising paradigm for efficient text-to-image synthesis. By distilling a latent consistency model (LCM) from a pre-trained teacher latent diffusion model (LDM), LCD facilitates the generation of high-fidelity images within merely 2 to 4 inference steps. However, the LCM's efficient inference is obtained at the cost of the sample quality. In this paper, we propose compensating the quality loss by aligning LCM's output with human preference during training. Specifically, we introduce Reward Guided LCD (RG-LCD), which integrates feedback from a reward model (RM) into the LCD process by augmenting the original LCD loss with the objective of maximizing the reward associated with LCM's single-step generation. As validated through human evaluation, when trained with the feedback of a good RM, the 2-step generations from our RG-LCM are favored by humans over the 50-step DDIM samples from the teacher LDM, representing a 25-time inference acceleration without quality loss. As directly optimizing towards differentiable RMs can suffer from over-optimization, we take the initial step to overcome this difficulty by proposing the use of a latent proxy RM (LRM). This novel component serves as an intermediary, connecting our LCM with the RM. Empirically, we demonstrate that incorporating the LRM into our RG-LCD successfully avoids high-frequency noise in the generated images, contributing to both improved Fréchet Inception Distance (FID) on MS-COCO and a higher HPSv2.1 score on HPSv2's test set, surpassing those achieved by the baseline LCM. Project Page: https://rg-lcd.github.io/

3234. On Bits and Bandits: Quantifying the Regret-Information Trade-off

链接：https://iclr.cc/virtual/2025/poster/31236 abstract： In many sequential decision problems, an agent performs a repeated task. He then suffers regret and obtains information that he may use in the following rounds. However, sometimes the agent may also obtain information and avoid suffering regret by querying external sources. We study the trade-off between the information an agent accumulates and the regret it suffers. We invoke information-theoretic methods for obtaining regret lower bounds, that also allow us to easily re-derive several known lower bounds. We introduce the first Bayesian regret lower bounds that depend on the information an agent accumulates. We also prove regret upper bounds using the amount of information the agent accumulates. These bounds show that information measured in bits, can be traded off for regret, measured in reward. Finally, we demonstrate the utility of these bounds in improving the performance of a question-answering task with large language models, allowing us to obtain valuable insights.

3235. HD-Painter: High-Resolution and Prompt-Faithful Text-Guided Image Inpainting with Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/30867 abstract： Recent progress in text-guided image inpainting, based on the unprecedented success of text-to-image diffusion models, has led to exceptionally realistic and visually plausible results. However, there is still significant potential for improvement in current text-to-image inpainting models, particularly in better aligning the inpainted area with user prompts. Therefore, we introduce $\textit{HD-Painter}$, a $\textbf{training-free}$ approach that $\textbf{accurately follows prompts}$. To this end, we design the $\textit{Prompt-Aware Introverted Attention (PAIntA)}$ layer enhancing self-attention scores by prompt information resulting in better text aligned generations. To further improve the prompt coherence we introduce the $\textit{Reweighting Attention Score Guidance (RASG)}$ mechanism seamlessly integrating a post-hoc sampling strategy into the general form of DDIM to prevent out-of-distribution latent shifts. Our experiments demonstrate that HD-Painter surpasses existing state-of-the-art approaches quantitatively and qualitatively across multiple metrics and a user study. Code is publicly available at: https://github.com/Picsart-AI-Research/HD-Painter

3236. Relax and Merge: A Simple Yet Effective Framework for Solving Fair k-Means and k-sparse Wasserstein Barycenter Problems

链接：https://iclr.cc/virtual/2025/poster/28424 abstract： The fairness of clustering algorithms has gained widespread attention across various areas, including machine learning, In this paper, we study fair k-means clustering in Euclidean space. Given a dataset comprising several groups, the fairness constraint requires that each cluster should contain a proportion of points from each group within specified lower and upper bounds. Due to these fairness constraints, determining the optimal locations of k centers is a quite challenging task. We propose a novel ``Relax and Merge'' framework that returns a $(1+4\rho + O(\epsilon))$-approximate solution, where ρ is the approximate ratio of an off-the-shelf vanilla k-means algorithm and $O(\epsilon)$ can be an arbitrarily small positive number. If equipped with a PTAS of k-means, our solution can achieve an approximation ratio of $(5+O(\epsilon))$ with only a slight violation of the fairness constraints, which improves the current state-of-the-art approximation guarantee. Furthermore, using our framework, we can also obtain a $(1+4\rho +O(\epsilon))$-approximate solution for the k-sparse Wasserstein Barycenter problem, which is a fundamental optimization problem in the field of optimal transport, and a $(2+6\rho)$-approximate solution for the strictly fair k-means clustering with no violation, both of which are better than the current state-of-the-art methods. In addition, the empirical results demonstrate that our proposed algorithm can significantly outperform baseline approaches in terms of clustering cost.

3237. JudgeBench: A Benchmark for Evaluating LLM-Based Judges

链接：https://iclr.cc/virtual/2025/poster/30311 abstract： LLM-based judges have emerged as a scalable alternative to human evaluation and are increasingly used to assess, compare, and improve models. However, the reliability of LLM-based judges themselves is rarely scrutinized. As LLMs become more advanced, their responses grow more sophisticated, requiring stronger judges to evaluate them. Existing benchmarks primarily focus on a judge's alignment with human preferences, but often fail to account for more challenging tasks where crowdsourced human preference is a poor indicator of factual and logical correctness. To address this, we propose a novel evaluation framework to objectively evaluate LLM-based judges. Based on this framework, we propose JudgeBench, a benchmark for evaluating LLM-based judges on challenging response pairs spanning knowledge, reasoning, math, and coding. JudgeBench leverages a novel pipeline for converting existing difficult datasets into challenging response pairs with preference labels reflecting objective correctness. Our comprehensive evaluation on a collection of prompted judges, fine-tuned judges, multi-agent judges, and reward models shows that JudgeBench poses a significantly greater challenge than previous benchmarks, with many strong models (e.g. GPT-4o) performing just slightly better than random guessing. Overall, JudgeBench offers a reliable platform for assessing increasingly advanced LLM-based judges. Data and code are available at \url{https://github.com/ScalerLab/JudgeBench}.

3238. SPaR: Self-Play with Tree-Search Refinement to Improve Instruction-Following in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30675 abstract： Instruction-following is a fundamental capability of language models, requiring the model to recognize even the most subtle requirements in the instructions and accurately reflect them in its output.Such an ability is well-suited for and often optimized by preference learning.However, existing methods often directly sample multiple independent responses from the model when creating preference pairs.Such practice can introduce content variations irrelevant to whether the instruction is precisely followed (e.g., different expressions about the same semantic), interfering with the goal of teaching models to recognize the key differences that lead to improved instruction following.In light of this, we introduce SPaR, a self-play framework integrating tree-search self-refinement to yield valid and comparable preference pairs free from distractions.By playing against itself, an LLM employs a tree-search strategy to refine its previous responses with respect to the instruction while minimizing unnecessary variations.Our experiments show that a LLaMA3-8B model, trained over three iterations guided by SPaR, surpasses GPT-4-Turbo on the IFEval benchmark without losing general capabilities. Furthermore, SPaR demonstrates promising scalability, greatly enhancing models like GLM-4-9B and LLaMA3-70B.We also identify how inference scaling in tree search would impact model performance.Our code and data are publicly available at https://github.com/thu-coai/SPaR.

3239. To Tackle Adversarial Transferability: A Novel Ensemble Training Method with Fourier Transformation

链接：https://iclr.cc/virtual/2025/poster/30054 abstract： Ensemble methods are commonly used for enhancing robustness in machine learning. However, due to the "transferability" of adversarial examples, the performance of an ensemble model can be seriously affected even it contains a set of independently trained sub-models. To address this issue, we propose an efficient data transformation method based on a cute "weakness allocation" strategy, to diversify non-robust features.Our approach relies on a fine-grained analysis on the relation between non-robust features and adversarial attack directions.Moreover, our approach enjoys several other advantages, e.g., it does not require any communication between sub-models and the construction complexity is also quite low.We conduct a set of experiments to evaluate the performance of our proposed method and compare it with several popular baselines. The results suggest that our approach can achieve significantly improved robust accuracy over most existing ensemble methods, and meanwhile preserve high clean accuracy.

3240. Boosting Methods for Interval-censored Data with Regression and Classification

链接：https://iclr.cc/virtual/2025/poster/30432 abstract： Boosting has garnered significant interest across both machine learning and statistical communities. Traditional boosting algorithms, designed for fully observed random samples, often struggle with real-world problems, particularly with interval-censored data. This type of data is common in survival analysis and time-to-event studies where exact event times are unobserved but fall within known intervals. Effective handling of such data is crucial in fields like medical research, reliability engineering, and social sciences. In this work, we introduce novel nonparametric boosting methods for regression and classification tasks with interval-censored data. Our approaches leverage censoring unbiased transformations to adjust loss functions and impute transformed responses while maintaining model accuracy. Implemented via functional gradient descent, these methods ensure scalability and adaptability. We rigorously establish their theoretical properties, including optimality and mean squared error trade-offs. Our proposed methods not only offer a robust framework for enhancing predictive accuracy in domains where interval-censored data are common but also complement existing work, expanding the applicability of existing boosting techniques. Empirical studies demonstrate robust performance across various finite-sample scenarios, highlighting the practical utility of our approaches.

3241. Multi-level Certified Defense Against Poisoning Attacks in Offline Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29326 abstract： Similar to other machine learning frameworks, Offline Reinforcement Learning (RL) is shown to be vulnerable to poisoning attacks, due to its reliance on externally sourced datasets, a vulnerability that is exacerbated by its sequential nature. To mitigate the risks posed by RL poisoning, we extend certified defenses to provide larger guarantees against adversarial manipulation, ensuring robustness for both per-state actions, and the overall expected cumulative reward. Our approach leverages properties of Differential Privacy, in a manner that allows this work to span both continuous and discrete spaces, as well as stochastic and deterministic environments---significantly expanding the scope and applicability of achievable guarantees. Empirical evaluations demonstrate that our approach ensures the performance drops to no more than 50% with up to 7% of the training data poisoned, significantly improving over the 0.008% in prior work (Wu et al., 2022), while producing certified radii that is 5 times larger as well. This highlights the potential of our framework to enhance safety and reliability in offline RL.

3242. GridMix: Exploring Spatial Modulation for Neural Fields in PDE Modeling

链接：https://iclr.cc/virtual/2025/poster/30317 abstract： Significant advancements have been achieved in PDE modeling using neural fields. Despite their effectiveness, existing methods rely on global modulation, limiting their ability to reconstruct local details. While spatial modulation with vanilla grid-based representations offers a promising alternative, it struggles with inadequate global information modeling and over-fitting to the training spatial domain. To address these challenges, we propose GridMix, a novel approach that models spatial modulation as a mixture of grid-based representations. GridMix effectively explores global structures while preserving locality for fine-grained modulation. Furthermore, we introduce spatial domain augmentation to enhance the robustness of the modulated neural fields against spatial domain variations. With all these innovations,our comprehensive approach culminates in MARBLE, a framework that significantly advancing the capabilities of neural fields in PDE modeling. The effectiveness of MARBLE is extensivelyvalidated on diverse benchmarks encompassing dynamics modeling and geometric prediction.

3243. Breaking Neural Network Scaling Laws with Modularity

链接：https://iclr.cc/virtual/2025/poster/30948 abstract： Modular neural networks outperform nonmodular neural networks on tasks ranging from visual question answering to robotics. These performance improvements are thought to be due to modular networks' superior ability to model the compositional and combinatorial structure of real-world problems. However, a theoretical explanation of how modularity improves generalizability, and how to leverage task modularity while training networks remains elusive. Using recent theoretical progress in explaining neural network generalization, we investigate how the amount of training data required to generalize on a task varies with the intrinsic dimensionality of a task's input. We show theoretically that when applied to modularly structured tasks, while nonmodular networks require an exponential number of samples with task dimensionality, modular networks' sample complexity is independent of task dimensionality: modular networks can generalize in high dimensions. We then develop a novel learning rule for modular networks to exploit this advantage and empirically show the improved generalization of the rule, both in- and out-of-distribution, on high-dimensional, modular tasks.

3244. Tractable Multi-Agent Reinforcement Learning through Behavioral Economics

链接：https://iclr.cc/virtual/2025/poster/28091 abstract： A significant roadblock to the development of principled multi-agent reinforcement learning (MARL) algorithms is the fact that desired solution concepts like Nash equilibria may be intractable to compute. We show how one can overcome this obstacle by introducing concepts from behavioral economics into MARL. To do so, we imbue agents with two key features of human decision-making: risk aversion and bounded rationality. We show that introducing these two properties into games gives rise to a class of equilibria---risk-averse quantal response equilibria (RQE)---which are tractable to compute in \emph{all} n-player matrix and finite-horizon Markov games. In particular, we show that they emerge as the endpoint of no-regret learning in suitably adjusted versions of the games. Crucially, the class of computationally tractable RQE is independent of the underlying game structure and only depends on agents' degrees of risk-aversion and bounded rationality. To validate the expressivity of this class of solution concepts we show that it captures peoples' patterns of play in a number of 2-player matrix games previously studied in experimental economics. Furthermore, we give a first analysis of the sample complexity of computing these equilibria in finite-horizon Markov games when one has access to a generative model. We validate our findings on a simple multi-agent reinforcement learning benchmark. Our results open the doors for to the principled development of new decentralized multi-agent reinforcement learning algorithms.

3245. ElasticTok: Adaptive Tokenization for Image and Video

链接：https://iclr.cc/virtual/2025/poster/28065 abstract： Efficient video tokenization remains a key bottleneck in learning general purpose vision models that are capable of processing long video sequences. Prevailing approaches are restricted to encoding videos to a fixed number of tokens, where too few tokens will result in overly lossy encodings, and too many tokens will result in prohibitively long sequence lengths. In this work, we introduce ElasticTok, a method that conditions on prior frames to adaptively encode a frame into a variable number of tokens. To enable this in a computationally scalable way, we propose a masking technique that drops a random number of tokens at the end of each frames's token encoding. During inference, ElasticTok can dynamically allocate tokens when needed -- more complex data can leverage more tokens, while simpler data only needs a few tokens. Our empirical evaluations on images and video demonstrate the effectiveness of our approach in efficient token usage, paving the way for future development of more powerful multimodal models, world models, and agents. Video examples of using ElasticTok can be found on our website: http://largeworldmodel.github.io/elastictok

3246. World Model on Million-Length Video And Language With Blockwise RingAttention

链接：https://iclr.cc/virtual/2025/poster/30229 abstract： Enabling long-context understanding remains a key challenge in scaling existing sequence models -- a crucial component in developing generally intelligent models that can process and operate over long temporal horizons that potentially consist of millions of tokens. In this paper, we aim to address these challenges by providing a comprehensive exploration of the full development process for producing 1M context language models and video-language models, setting new benchmarks in language retrieval and new capabilities in long video understanding. We detail our long context data curation process, progressive context extension from 4K to 1M tokens, and present an efficient open-source implementation for scalable training on long sequences. Additionally, we open-source a family of 7B parameter models capable of processing long text documents and videos exceeding 1M tokens.

3247. RocketEval: Efficient automated LLM evaluation via grading checklist

链接：https://iclr.cc/virtual/2025/poster/27674 abstract： Evaluating large language models (LLMs) in diverse and challenging scenarios is essential to align them with human preferences. To mitigate the prohibitive costs associated with human evaluations, utilizing a powerful LLM as a judge has emerged as a favored approach. Nevertheless, this methodology encounters several challenges, including substantial expenses, concerns regarding privacy and security, and reproducibility. In this paper, we propose a straightforward, replicable, and accurate automated evaluation method by leveraging a lightweight LLM as the judge, named RocketEval. Initially, we identify that the performance disparity between lightweight and powerful LLMs in evaluation tasks primarily stems from their ability to conduct comprehensive analyses, which is not easily enhanced through techniques such as chain-of-thought reasoning. By reframing the evaluation task as a multi-faceted Q\&A using an instance-specific checklist, we demonstrate that the limited judgment accuracy of lightweight LLMs is largely attributes to high uncertainty and positional bias. To address these challenges, we introduce an automated evaluation process grounded in checklist grading, which is designed to accommodate a variety of scenarios and questions. This process encompasses the creation of checklists, the grading of these checklists by lightweight LLMs, and the reweighting of checklist items to align with the supervised annotations. Our experiments carried out on the automated evaluation benchmarks, MT-Bench and WildBench datasets, reveal that RocketEval, when using $\textit{Gemma-2-2B}$ as the judge, achieves a high correlation (0.965) with human preferences, which is comparable to $\textit{GPT-4o}$. Moreover, RocketEval provides a cost reduction exceeding 50-fold for large-scale evaluation and comparison scenarios. Our code is available at https://github.com/Joinn99/RocketEval-ICLR.

3248. Look Before You Leap: Universal Emergent Mechanism for Retrieval in Language Models

链接：https://iclr.cc/virtual/2025/poster/28935 abstract： When solving challenging problems, language models (LMs) are able to identify relevant information from long and complicated contexts. To study how LMs solve retrieval tasks in diverse situations, we introduce ORION, a collection of structured retrieval tasks spanning six domains, from text understanding to coding. Each task in ORION can be represented abstractly by a request (e.g. a question) that retrieves an attribute (e.g. the character name) from a context (e.g. a story). We apply causal analysis on 18 open-source language models with sizes ranging from 125 million to 70 billion parameters. We find that LMs internally decompose retrieval tasks in a modular way: middle layers at the last token position process the request, while late layers retrieve the correct entity from the context. After causally enforcing this decomposition, models are still able to solve the original task, preserving 70% of the original correct token probability in 98 of the 106 studied model-task pairs. We connect our macroscopic decomposition with a microscopic description by performing a fine-grained case study of a question-answering task on Pythia-2.8b. Building on our high-level understanding, we demonstrate a proof of concept application for scalable internal oversight of LMs to mitigate prompt-injection while requiring human supervision on only a single input. Our solution improves accuracy drastically (from 15.5% to 97.5% on Pythia-12b). This work presents evidence of a universal emergent modular processing of tasks across varied domains and models and is a pioneering effort in applying interpretability for scalable internal oversight of LMs.

3249. The Hidden Cost of Waiting for Accurate Predictions

链接：https://iclr.cc/virtual/2025/poster/30653 abstract： Algorithmic predictions are increasingly informing societal resource allocations by identifying individuals for targeting. Policymakers often build these systems with the assumption that by gathering more observations on individuals, they can improve predictive accuracy and, consequently, allocation efficiency. An overlooked yet consequential aspect of prediction-driven allocations is that of timing. The planner has to trade off relying on earlier and potentially noisier predictions to intervene before individuals experience undesirable outcomes, or they may wait to gather more observations to make more precise allocations. We examine this tension using a simple mathematical model, where the planner collects observations on individuals to improve predictions over time. We analyze both the ranking induced by these predictions and optimal resource allocation. We show that though individual prediction accuracy improves over time, counter-intuitively, the average ranking loss can worsen. As a result, the planner's ability to improve social welfare can decline. We identify inequality as a driving factor behind this phenomenon. Our findings provide a nuanced perspective and challenge the conventional wisdom that it is preferable to wait for more accurate predictions to ensure the most efficient allocations.

3250. Boosting Latent Diffusion with Perceptual Objectives

链接：https://iclr.cc/virtual/2025/poster/27753 abstract： Latent diffusion models (LDMs) power state-of-the-art high-resolution generative image models. LDMs learn the data distribution in the latent space of an autoencoder (AE) and produce images by mapping the generated latents into RGB image space using the AE decoder. While this approach allows for efficient model training and sampling, it induces a disconnect between the training of the diffusion model and the decoder, resulting in a loss of detail in the generated images. To remediate this disconnect, we propose to leverage the internal features of the decoder to define a latent perceptual loss (LPL). This loss encourages the models to create sharper and more realistic images. Our loss can be seamlessly integrated with common autoencoders used in latent diffusion models, and can be applied to different generative modeling paradigms such as DDPM with epsilon and velocity prediction, as well as flow matching. Extensive experiments with models trained on three datasets at 256 and 512 resolution show improved quantitative -- with boosts between 6% and 20% in FID -- and qualitative results when using our perceptual loss.

3251. Do LLMs ``know'' internally when they follow instructions?

链接：https://iclr.cc/virtual/2025/poster/28257 abstract： Instruction-following is crucial for building AI agents with large language models (LLMs), as these models must adhere strictly to user-provided constraints and guidelines. However, LLMs often fail to follow even simple and clear instructions.To improve instruction-following behavior and prevent undesirable outputs, a deeper understanding of how LLMs' internal states relate to these outcomes is required.In this work, we investigate whether LLMs encode information in their representations that correlates with instruction-following success—a property we term ``knowing internally''.Our analysis identifies a direction in the input embedding space, termed the instruction-following dimension, that predicts whether a response will comply with a given instruction.We find that this dimension generalizes well across unseen tasks but not across unseen instruction types.We demonstrate that modifying representations along this dimension improves instruction-following success rates compared to random changes, without compromising response quality.Further investigation reveals that this dimension is more closely related to the phrasing of prompts rather than the inherent difficulty of the task or instructions. This discovery also suggests explanations for why LLMs sometimes fail to follow clear instructions and why prompt engineering is often effective, even when the content remains largely unchanged. This work provides insight into the internal workings of LLMs' instruction-following, paving the way for reliable LLM agents.

3252. RelCon: Relative Contrastive Learning for a Motion Foundation Model for Wearable Data

链接：https://iclr.cc/virtual/2025/poster/28603 abstract： We present RelCon, a novel self-supervised Relative Contrastive learning approach for training a motion foundation model from wearable accelerometry sensors. First, a learnable distance measure is trained to capture motif similarity and domain-specific semantic information such as rotation invariance. Then, the learned distance provides a measurement of semantic similarity between a pair of accelerometry time-series, which we use to train our foundation model to model relative relationships across time and across subjects. The foundation model is trained on 1 billion segments from 87,376 participants, and achieves strong performance across multiple downstream tasks, including human activity recognition and gait metric regression. To our knowledge, we are the first to show the generalizability of a foundation model with motion data from wearables across distinct evaluation tasks.

3253. EFFICIENT JAILBREAK ATTACK SEQUENCES ON LARGE LANGUAGE MODELS VIA MULTI-ARMED BANDIT-BASED CONTEXT SWITCHING

链接：https://iclr.cc/virtual/2025/poster/28647 abstract： Content warning: This paper contains examples of harmful language and content.Recent advances in large language models (LLMs) have made them increasingly vulnerable to jailbreaking attempts, where malicious users manipulate models into generating harmful content. While existing approaches rely on either single-step attacks that trigger immediate safety responses or multi-step methods that inefficiently iterate prompts using other LLMs, we introduce ``Sequence of Context'' (SoC) attacks that systematically alter conversational context through strategically crafted context-switching queries (CSQs). We formulate this as a multi-armed bandit (MAB) optimization problem, automatically learning optimal sequences of CSQs that gradually weaken the model's safety boundaries. Our theoretical analysis provides tight bounds on both the expected sequence length until successful jailbreak and the convergence of cumulative rewards. Empirically, our method achieves a 95\% attack success rate, surpassing PAIR by 63.15\%, AutoDAN by 60\%, and ReNeLLM by 50\%. We evaluate our attack across multiple open-source LLMs including Llama and Mistral variants. Our findings highlight critical vulnerabilities in current LLM safeguards and emphasize the need for defenses that consider sequential attack patterns rather than relying solely on static prompt filtering or iterative refinement.

3254. Diff3DS: Generating View-Consistent 3D Sketch via Differentiable Curve Rendering

链接：https://iclr.cc/virtual/2025/poster/29181 abstract： 3D sketches are widely used for visually representing the 3D shape and structure of objects or scenes. However, the creation of 3D sketch often requires users to possess professional artistic skills. Existing research efforts primarily focus on enhancing the ability of interactive sketch generation in 3D virtual systems. In this work, we propose Diff3DS, a novel differentiable rendering framework for generating view-consistent 3D sketch by optimizing 3D parametric curves under various supervisions. Specifically, we perform perspective projection to render the 3D rational Bézier curves into 2D curves, which are subsequently converted to a 2D raster image via our customized differentiable rasterizer. Our framework bridges the domains of 3D sketch and raster image, achieving end-to-end optimization of 3D sketch through gradients computed in the 2D image domain. Our Diff3DS can enable a series of novel 3D sketch generation tasks, including text-to-3D sketch and image-to-3D sketch, supported by the popular distillation-based supervision, such as Score Distillation Sampling (SDS). Extensive experiments have yielded promising results and demonstrated the potential of our framework. Project: https://yiboz2001.github.io/Diff3DS/

3255. Adaptive Length Image Tokenization via Recurrent Allocation

链接：https://iclr.cc/virtual/2025/poster/28458 abstract： Current vision systems typically assign fixed-length representations to images, regardless of the information content. This contrasts with human intelligence —and even large language models—which allocate varying representational capacities based on entropy, context and familiarity. Inspired by this, we propose an approach to learn variable-length token representations for 2D images. Our encoder-decoder architecture recursively processes 2D image tokens, distilling them into 1D latent tokens over multiple iterations of recurrent rollouts. Each iteration refines the 2D tokens, updates the existing 1D latent tokens, and adaptively increases representational capacity by adding new tokens. This enables compression of images into a variable number of tokens, ranging from 32 to 256. We validate our tokenizer using reconstruction loss and FID metrics, demonstrating that token count aligns with image entropy, familiarity and downstream task requirements. Recurrent token processing with increasing representational capacity in each iteration shows signs of token specialization, revealing potential for object / part discovery.

3256. Diffusing States and Matching Scores: A New Framework for Imitation Learning

链接：https://iclr.cc/virtual/2025/poster/28575 abstract： Adversarial Imitation Learning is traditionally framed as a two-player zero-sum game between a learner and an adversarially chosen cost function, and can therefore be thought of as the sequential generalization of a Generative Adversarial Network (GAN). However, in recent years, diffusion models have emerged as a non-adversarial alternative to GANs that merely require training a score function via regression, yet produce generations of higher quality. In response, we investigate how to lift insights from diffusion modeling to the sequential setting. We propose diffusing states and performing score-matching along diffused states to measure the discrepancy between the expert's and learner's states. Thus, our approach only requires training score functions to predict noises via standard regression, making it significantly easier and more stable to train than adversarial methods. Theoretically, we prove first- and second-order instance-dependent bounds with linear scaling in the horizon, proving that our approach avoids the compounding errors that stymie offline approaches to imitation learning. Empirically, we show our approach outperforms both GAN-style imitation learning baselines and discriminator-free imitation learning baselines across various continuous control problems, including complex tasks like controlling humanoids to walk, sit, crawl, and navigate through obstacles.

3257. Correlated Proxies: A New Definition and Improved Mitigation for Reward Hacking

链接：https://iclr.cc/virtual/2025/poster/28442 abstract： Because it is difficult to precisely specify complex objectives, reinforcement learning policies are often optimized using proxy reward functions that only approximate the true goal. However, optimizing proxy rewards frequently leads to reward hacking: the optimized reward function ceases to be a good proxy and the resulting policy performs poorly with respect to the unspecified true reward. Principled solutions to reward hacking have been impeded by the lack of a good definition for the problem. To address this gap, we introduce a definition of reward hacking based on the correlation between proxy and true rewards for states and actions seen by a “reference policy” that breaks down under optimization. We show that this definition captures reward hacking behavior across several realistic settings, including in reinforcement learning from human feedback (RLHF). Using our formulation, we show theoretically that regularization to the reference policy can effectively prevent reward hacking. While the current practice in RLHF applies a KL penalty between action distributions for this purpose, our theory suggests regularizing the χ2 divergence between the policies’ occupancy measures can be more effective. We intuitively show the benefits of this type of regularization and demonstrate that it better mitigates reward hacking in practice across four realistic settings, including RLHF. Our code is available at https://github.com/cassidylaidlaw/orpo.

3258. Positional Embeddings in Transformer Models: Evolution from Text to Vision Domains

链接：https://iclr.cc/virtual/2025/poster/31345 abstract： Positional encoding has become an essential element in transformer models, addressing their fundamental property of permutation invariance and allowing them to understand sequential relationships within data. This blog post examines positional encoding techniques, emphasizing their vital importance in traditional transformers and their use with 2D data in Vision Transformers (ViT). We explore two contemporary methods—ALiBi (Attention with Linear Biases) and RoPE (Rotary Position Embedding)—analyzing their unique approaches to tackling the challenge of sequence length extrapolation during inference, a significant issue for transformers. Additionally, we compare these methods' fundamental similarities and differences, assessing their impact on transformer performance across various fields. We also look into how interpolation strategies have been utilized to enhance the extrapolation capabilities of these methods; we conclude this blog with an empirical comparison of ALiBi and RoPE in Vision Transformers. To the best of our knowledge, this represents the first direct comparison of these positional encoding methods with those used in standard Vision Transformers.

3259. Natural Language Inference Improves Compositionality in Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/30305 abstract： Compositional reasoning in Vision-Language Models (VLMs) remains challenging as these models often struggle to relate objects, attributes, and spatial relationships. Recent methods aim to address these limitations by relying on the semantics of the textual description, using Large Language Models (LLMs) to break them down into subsets of questions and answers. However, these methods primarily operate on the surface level, failing to incorporate deeper lexical understanding while introducing incorrect assumptions generated by the LLM. In response to these issues, we present Caption Expansion with Contradictions and Entailments (CECE), a principled approach that leverages Natural Language Inference (NLI) to generate entailments and contradictions from a given premise. CECE produces lexically diverse sentences while maintaining their core meaning. Through extensive experiments, we show that CECE enhances interpretability and reduces overreliance on biased or superficial features. By balancing CECE along the original premise, we achieve significant improvements over previous methods without requiring additional fine-tuning, producing state-of-the-art results on benchmarks that score agreement with human judgments for image-text alignment, and achieving an increase in performance on Winoground of $+19.2\%$ (group score) and $+12.9\%$ on EqBen (group score) over the best prior work (finetuned with targeted data).

3260. Language models scale reliably with over-training and on downstream tasks

链接：https://iclr.cc/virtual/2025/poster/28688 abstract： Scaling laws are useful guides for derisking expensive training runs, as they predict performance of large models using cheaper, small-scale experiments. However, there remain gaps between current scaling studies and how language models are ultimately trained and evaluated. For instance, scaling is usually studied in the compute-optimal training regime (i.e., "Chinchilla optimal" regime). In contrast, models are often over-trained to reduce inference costs. Moreover, scaling laws mostly predict loss on next-token prediction, but models are usually compared on downstream task performance. To address both shortcomings, we create a testbed of 104 models with 0.011B to 6.9B parameters trained with various numbers of tokens on three data distributions. First, we fit scaling laws that extrapolate in both the amount of over-training and the number of model parameters. This enables us to predict the validation loss of a 1.4B parameter, 900B token run (i.e., 32$\times$ over-trained) and a 6.9B parameter, 138B token run (i.e., a compute-optimal run)—each from experiments that take 300$\times$ less compute. Second, we relate the perplexity of a language model to its downstream task performance by proposing a power law. We use this law to predict top-1 error averaged over downstream tasks for the two aforementioned models, using experiments that take 20$\times$ less compute.

3261. MMR: A Large-scale Benchmark Dataset for Multi-target and Multi-granularity Reasoning Segmentation

链接：https://iclr.cc/virtual/2025/poster/28436 abstract： The fusion of Large Language Models (LLMs) with vision models is pioneering new possibilities in user-interactive vision-language tasks. A notable application is reasoning segmentation, where models generate pixel-level segmentation masks by comprehending implicit meanings in human instructions. However, seamless human-AI interaction demands more than just object-level recognition; it requires understanding both objects and the functions of their detailed parts, particularly in multi-target scenarios. For example, when instructing a robot to \textit{“turn on the TV"}, there could be various ways to accomplish this command. Recognizing multiple objects capable of turning on the TV, such as the TV itself or a remote control (multi-target), provides more flexible options and aids in finding the optimized scenario. Furthermore, understanding specific parts of these objects, like the TV's button or the remote's button (part-level), is important for completing the action. Unfortunately, current reasoning segmentation datasets predominantly focus on a single target object-level reasoning, which limits the detailed recognition of an object's parts in multi-target contexts. To address this gap, we construct a large-scale dataset called Multi-target and Multi-granularity Reasoning (MMR). MMR comprises 194K complex and implicit instructions that consider multi-target, object-level, and part-level aspects, based on pre-existing image-mask sets. This dataset supports diverse and context-aware interactions by hierarchically providing object and part information. Moreover, we propose a straightforward yet effective framework for multi-target, object-level, and part-level reasoning segmentation. Experimental results on MMR show that the proposed method can reason effectively in multi-target and multi-granularity scenarios, while the existing reasoning segmentation model still has room for improvement. The dataset is available at \url{https://github.com/jdg900/MMR}.

3262. Diffusion-based Decoupled Deterministic and Uncertain Framework for Probabilistic Multivariate Time Series Forecasting

链接：https://iclr.cc/virtual/2025/poster/30216 abstract： Diffusion-based denoising models have demonstrated impressive performance in probabilistic forecasting for multivariate time series (MTS). Nonetheless, existing approaches often model the entire data distribution, neglecting the variability in uncertainty across different components of the time series. This paper introduces a Diffusion-based Decoupled Deterministic and Uncertain ($\mathrm{D^3U}$) framework for probabilistic MTS forecasting. The framework integrates non-probabilistic forecasting with conditional diffusion generation, enabling both accurate point predictions and probabilistic forecasting. $\mathrm{D^3U}$ utilizes a point forecasting model to non-probabilistically model high-certainty components in the time series, generating embedded representations that are conditionally injected into a diffusion model. To better model high-uncertainty components, a patch-based denoising network (PatchDN) is designed in the conditional diffusion model. Designed as a plug-and-play framework, $\mathrm{D^3U}$ can be seamlessly integrated into existing point forecasting models to provide probabilistic forecasting capabilities. It can also be applied to other conditional diffusion methods that incorporate point forecasting models. Experiments on six real-world datasets demonstrate that our method achieves over a 20\% improvement in both point and probabilistic forecasting performance in MTS long-term forecasting compared to state-of-the-art (SOTA) probabilistic forecasting methods. Additionally, extensive ablation studies further validate the effectiveness of the $\mathrm{D^3U}$ framework.

3263. Episodic Novelty Through Temporal Distance

链接：https://iclr.cc/virtual/2025/poster/30190 abstract： Exploration in sparse reward environments remains a significant challenge in reinforcement learning, particularly in Contextual Markov Decision Processes (CMDPs), where environments differ across episodes. Existing episodic intrinsic motivation methods for CMDPs primarily rely on count-based approaches, which are ineffective in large state spaces, or on similarity-based methods that lack appropriate metrics for state comparison. To address these shortcomings, we propose Episodic Novelty Through Temporal Distance (ETD), a novel approach that introduces temporal distance as a robust metric for state similarity and intrinsic reward computation. By employing contrastive learning, ETD accurately estimates temporal distances and derives intrinsic rewards based on the novelty of states within the current episode. Extensive experiments on various benchmark tasks demonstrate that ETD significantly outperforms state-of-the-art methods, highlighting its effectiveness in enhancing exploration in sparse reward CMDPs.

3264. Mix-LN: Unleashing the Power of Deeper Layers by Combining Pre-LN and Post-LN

链接：https://iclr.cc/virtual/2025/poster/30583 abstract： Large Language Models (LLMs) have achieved remarkable success, yet recent findings reveal that their deeper layers often contribute minimally and can be pruned without affecting overall performance. While some view this as an opportunity for model compression, we identify it as a training shortfall rooted in the widespread use of Pre-Layer Normalization (Pre-LN). We demonstrate that Pre-LN, commonly employed in models like GPT and LLaMA, leads to diminished gradient norms in its deeper layers, reducing their effectiveness. In contrast, Post-Layer Normalization (Post-LN) preserves larger gradient norms in deeper layers but suffers from vanishing gradients in earlier layers. To address this, we introduce Mix-LN, a novel normalization technique that combines the strengths of Pre-LN and Post-LN within the same model. Mix-LN applies Post-LN to the earlier layers and Pre-LN to the deeper layers, ensuring more uniform gradient norms across layers. This allows all parts of the network—both shallow and deep layers—to contribute effectively to training. Extensive experiments with various model sizes demonstrate that Mix-LN consistently outperforms both Pre-LN and Post-LN, promoting more balanced, healthier gradient norms throughout the network, and enhancing the overall quality of LLM pre-training. Furthermore, we demonstrate that models pre-trained with Mix-LN learn better compared to those using Pre-LN or Post-LN during supervised fine-tuning, highlighting the critical importance of high-quality deep layers. By effectively addressing the inefficiencies of deep layers in current LLMs, Mix-LN unlocks their potential, enhancing model capacity without increasing model size. Our code is available at https://github.com/pixeli99/MixLN.

3265. VICtoR: Learning Hierarchical Vision-Instruction Correlation Rewards for Long-horizon Manipulation

链接：https://iclr.cc/virtual/2025/poster/29454 abstract： We study reward models for long-horizon manipulation by learning from action-free videos and language instructions, which we term the visual-instruction correlation (VIC) problem. Existing VIC methods face challenges in learning rewards for long-horizon tasks due to their lack of sub-stage awareness, difficulty in modeling task complexities, and inadequate object state estimation. To address these challenges,we introduce VICtoR, a novel hierarchical VIC reward model capable of providing effective reward signals for long-horizon manipulation tasks. Trained solely on primitive motion demonstrations, VICtoR effectively provides precise reward signals for long-horizon tasks by assessing task progress at various stages using a novel stage detector and motion progress evaluator. We conducted extensive experiments in both simulated and real-world datasets. The results suggest that VICtoR outperformed the best existing methods, achieving a 43% improvement in success rates for long-horizon tasks. Our project page can be found at https://cmlab-victor.github.io/cmlab-vicotor.github.io/.

3266. Fewer May Be Better: Enhancing Offline Reinforcement Learning with Reduced Dataset

链接：https://iclr.cc/virtual/2025/poster/27646 abstract： Research in offline reinforcement learning (RL) marks a paradigm shift in RL. However, a critical yet under-investigated aspect of offline RL is determining the subset of the offline dataset, which is used to improve algorithm performance while accelerating algorithm training. Moreover, the size of reduced datasets can uncover the requisite offline data volume essential for addressing analogous challenges. Based on the above considerations, we propose identifying Reduced Datasets for Offline RL (ReDOR) by formulating it as a gradient approximation optimization problem. We prove that the common actor-critic framework in reinforcement learning can be transformed into a submodular objective. This insight enables us to construct a subset by adopting the orthogonal matching pursuit (OMP). Specifically, we have made several critical modifications to OMP to enable successful adaptation with Offline RL algorithms. The experimental results indicate that the data subsets constructed by the ReDOR can significantly improve algorithm performance with low computational complexity.

3267. Context-Alignment: Activating and Enhancing LLMs Capabilities in Time Series

链接：https://iclr.cc/virtual/2025/poster/28084 abstract： Recently, leveraging pre-trained Large Language Models (LLMs) for time series (TS) tasks has gained increasing attention, which involves activating and enhancing LLMs' capabilities. Many methods aim to activate LLMs' capabilities based on token-level alignment, but overlook LLMs' inherent strength in natural language processing — their deep understanding of linguistic logic and structure rather than superficial embedding processing. We propose Context-Alignment (CA), a new paradigm that aligns TS with a linguistic component in the language environments familiar to LLMs to enable LLMs to contextualize and comprehend TS data, thereby activating their capabilities. Specifically, such context-level alignment comprises structural alignment and logical alignment, which is achieved by Dual-Scale Context-Alignment GNNs (DSCA-GNNs) applied to TS-language multimodal inputs. Structural alignment utilizes dual-scale nodes to describe hierarchical structure in TS-language, enabling LLMs to treat long TS data as a whole linguistic component while preserving intrinsic token features. Logical alignment uses directed edges to guide logical relationships, ensuring coherence in the contextual semantics. Following the DSCA-GNNs framework, we propose an instantiation method of CA, termed Few-Shot prompting Context-Alignment (FSCA), to enhance the capabilities of pre-trained LLMs in handling TS tasks. FSCA can be flexibly and repeatedly integrated into various layers of pre-trained LLMs to improve awareness of logic and structure, thereby enhancing performance. Extensive experiments show the effectiveness of FSCA and the importance of Context-Alignment across tasks, particularly in few-shot and zero-shot forecasting, confirming that Context-Alignment provides powerful prior knowledge on context. The code is open-sourced at https://github.com/tokaka22/ICLR25-FSCA.

3268. ECD: A Machine Learning Benchmark for Predicting Enhanced-Precision Electronic Charge Density in Crystalline Inorganic Materials

链接：https://iclr.cc/virtual/2025/poster/29623 abstract： Supervised machine learning techniques are increasingly being adopted to speed up electronic structure predictions, serving as alternatives to first-principles methods like Density Functional Theory (DFT). Although current DFT datasets mainly emphasize chemical properties and atomic forces, the precise prediction of electronic charge density is essential for accurately determining a system's total energy and ground state properties. In this study, we introduce a novel electronic charge density dataset named ECD, which encompasses 140,646 stable crystal geometries with medium-precision Perdew–Burke–Ernzerhof (PBE) functional data. Within this dataset, a subset of 7,147 geometries includes high-precision electronic charge density data calculated using the Heyd–Scuseria–Ernzerhof (HSE) functional in DFT. By designing various benchmark tasks for crystalline materials and emphasizing training with large-scale PBE data while fine-tuning with a smaller subset of high-precision HSE data, we demonstrate the efficacy of current machine learning models in predicting electronic charge densities.The ECD dataset and baseline models are open-sourced to support community efforts in developing new methodologies and accelerating materials design and applications.

3269. Harnessing Diversity for Important Data Selection in Pretraining Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29114 abstract： Data selection is of great significance in pretraining large language models, given the variation in quality within the large-scale available training corpora. To achieve this, researchers are currently investigating the use of data influence to measure the importance of data instances, $i.e.,$ a high influence score indicates that incorporating this instance to the training set is likely to enhance the model performance. Consequently, they select the top-k instances with the highest scores. However, this approach has several limitations. (1) Calculating the accurate influence of all available data is time-consuming.(2) The selected data instances are not diverse enough, which may hinder the pretrained model's ability to generalize effectively to various downstream tasks.In this paper, we introduce $\texttt{Quad}$, a data selection approach that considers both quality and diversity by using data influence to achieve state-of-the-art pretraining results.To compute the influence ($i.e.,$ the quality) more accurately and efficiently, we incorporate the attention layers to capture more semantic details, which can be accelerated through the Kronecker product. For the diversity, $\texttt{Quad}$ clusters the dataset into similar data instances within each cluster and diverse instances across different clusters. For each cluster, if we opt to select data from it, we take some samples to evaluate the influence to prevent processing all instances. Overall, we favor clusters with highly influential instances (ensuring high quality) or clusters that have been selected less frequently (ensuring diversity), thereby well balancing between quality and diversity. Experiments on Slimpajama and FineWeb over 7B large language models demonstrate that $\texttt{Quad}$ significantly outperforms other data selection methods with a low FLOPs consumption. Further analysis also validates the effectiveness of our influence calculation.

3270. Global Well-posedness and Convergence Analysis of Score-based Generative Models via Sharp Lipschitz Estimates

链接：https://iclr.cc/virtual/2025/poster/28221 abstract： We establish global well-posedness and convergence of the score-based generative models (SGM) under minimal general assumptions of initial data for score estimation. For the smooth case, we start from a Lipschitz bound of the score function with optimal time length. The optimality is validated by an example whose Lipschitz constant of scores is bounded at initial but blows up in finite time. This necessitates the separation of time scales in conventional bounds for non-log-concave distributions. In contrast, our follow up analysis only relies on a local Lipschitz condition and is valid globally in time. This leads to the convergence of numerical scheme without time separation. For the non-smooth case, we show that the optimal Lipschitz bound is $O(1/t)$ in the point-wise sense for distributions supported on a compact, smooth and low-dimensional manifold with boundary.

3271. Domain Guidance: A Simple Transfer Approach for a Pre-trained Diffusion Model

链接：https://iclr.cc/virtual/2025/poster/29727 abstract： Recent advancements in diffusion models have revolutionized generative modeling. However, the impressive and vivid outputs they produce often come at the cost of significant model scaling and increased computational demands. Consequently, building personalized diffusion models based on off-the-shelf models has emerged as an appealing alternative. In this paper, we introduce a novel perspective on conditional generation for transferring a pre-trained model. From this viewpoint, we propose *Domain Guidance*, a straightforward transfer approach that leverages pre-trained knowledge to guide the sampling process toward the target domain. Domain Guidance shares a formulation similar to advanced classifier-free guidance, facilitating better domain alignment and higher-quality generations. We provide both empirical and theoretical analyses of the mechanisms behind Domain Guidance. Our experimental results demonstrate its substantial effectiveness across various transfer benchmarks, achieving over a 19.6\% improvement in FID and a 23.4\% improvement in FD$_\text{DINOv2}$ compared to standard fine-tuning. Notably, existing fine-tuned models can seamlessly integrate Domain Guidance to leverage these benefits, without additional training.

3272. VisualPredicator: Learning Abstract World Models with Neuro-Symbolic Predicates for Robot Planning

链接：https://iclr.cc/virtual/2025/poster/29691 abstract： Broadly intelligent agents should form task-specific abstractions that selectively expose the essential elements of a task, while abstracting away the complexity of the raw sensorimotor space. In this work, we present Neuro-Symbolic Predicates, a first-order abstraction language that combines the strengths of symbolic and neural knowledge representations. We outline an online algorithm for inventing such predicates and learning abstract world models. We compare our approach to hierarchical reinforcement learning, vision-language model planning, and symbolic predicate invention approaches, on both in- and out-of-distribution tasks across five simulated robotic domains. Results show that our approach offers better sample complexity, stronger out-of-distribution generalization, and improved interpretability.

3273. Linear Multistep Solver Distillation for Fast Sampling of Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/27893 abstract： Sampling from diffusion models can be seen as solving the corresponding probability flow ordinary differential equation (ODE). The solving process requires a significant number of function evaluations (NFE), making it time-consuming. Recently, several solver search frameworks have attempted to find better-performing model-specific solvers. However, predicting the impact of intermediate solving strategies on final sample quality remains challenging, rendering the search process inefficient. In this paper, we propose a novel method for designing solving strategies. We first introduce a unified prediction formula for linear multistep solvers. Subsequently, we present a solver distillation framework, which enables a student solver to mimic the sampling trajectory generated by a teacher solver with more steps. We utilize the mean Euclidean distance between the student and teacher sampling trajectories as a metric, facilitating rapid adjustment and optimization of intermediate solving strategies. The design space of our framework encompasses multiple aspects, including prediction coefficients, time step schedules, and time scaling factors. Our framework has the ability to complete a solver search for Stable-Diffusion in under 12 total GPU hours. Compared to previous reinforcement learning-based search frameworks, our approach achieves over a 10$\times$ increase in search efficiency. With just 5 NFE, we achieve FID scores of 3.23 on CIFAR10, 7.16 on ImageNet-64, 5.44 on LSUN-Bedroom, and 12.52 on MS-COCO, resulting in a 2$\times$ sampling acceleration ratio compared to handcrafted solvers.

3274. BodyGen: Advancing Towards Efficient Embodiment Co-Design

链接：https://iclr.cc/virtual/2025/poster/29047 abstract： Embodiment co-design aims to optimize a robot's morphology and control policy simultaneously. While prior work has demonstrated its potential for generating environment-adaptive robots, this field still faces persistent challenges in optimization efficiency due to the (i) combinatorial nature of morphological search spaces and (ii) intricate dependencies between morphology and control.We prove that the ineffective morphology representation and unbalanced reward signals between the design and control stages are key obstacles to efficiency.To advance towards efficient embodiment co-design, we propose BodyGen, which utilizes (1) topology-aware self-attention for both design and control, enabling efficient morphology representation with lightweight model sizes; (2) a temporal credit assignment mechanism that ensures balanced reward signals for optimization. With our findings, BodyGen achieves an average 60.03% performance improvement against state-of-the-art baselines. We provide codes and more results on the website: https://genesisorigin.github.io.

3275. ProtoSnap: Prototype Alignment For Cuneiform Signs

链接：https://iclr.cc/virtual/2025/poster/29313 abstract： The cuneiform writing system served as the medium for transmitting knowledgein the ancient Near East for a period of over three thousand years. Cuneiformsigns have a complex internal structure which is the subject of expert paleographicanalysis, as variations in sign shapes bear witness to historical developments andtransmission of writing and culture over time. However, prior automated techniquesmostly treat sign types as categorical and do not explicitly model their highly variedinternal configurations. In this work, we present an unsupervised approach forrecovering the fine-grained internal configuration of cuneiform signs by leveragingpowerful generative models and the appearance and structure of prototype fontimages as priors. Our approach, ProtoSnap, enforces structural consistency onmatches found with deep image features to estimate the diverse configurationsof cuneiform characters, snapping a skeleton-based template to photographedcuneiform signs. We provide a new benchmark of expert annotations and evaluateour method on this task. Our evaluation shows that our approach succeeds inaligning prototype skeletons to a wide variety of cuneiform signs. Moreover, weshow that conditioning on structures produced by our method allows for generatingsynthetic data with correct structural configurations, significantly boosting theperformance of cuneiform sign recognition beyond existing techniques, in particularover rare signs. Our code, data, and trained models are available at the project page:https://tau-vailab.github.io/ProtoSnap/

3276. Weighted-Reward Preference Optimization for Implicit Model Fusion

链接：https://iclr.cc/virtual/2025/poster/28854 abstract： While fusing heterogeneous open-source LLMs with varying architectures and sizes can potentially integrate the strengths of different models, existing fusion methods face significant challenges, such as vocabulary alignment and merging distribution matrices. These procedures are not only complex but also prone to introducing noise and errors. In this paper, we propose an implicit fusion method, Weighted-Reward Preference Optimization (WRPO), which leverages preference optimization between the source LLMs and the target LLM to transfer their capabilities effectively. WRPO eliminates the need for vocabulary alignment and matrix fusion and can be efficiently scaled to accommodate various LLMs. To address distributional deviations between the source and target LLMs, WRPO introduces a progressive adaptation strategy that gradually shifts reliance on preferred examples from the target LLM to the source LLMs. Extensive experiments on the MT-Bench, AlpacaEval-2, and Arena-Hard benchmarks demonstrate that WRPO consistently outperforms existing knowledge fusion methods and various fine-tuning baselines. When applied to LLaMA3-8B-Instruct as the target model, WRPO achieves a length-controlled win rate of 55.9\% against GPT-4-Preview-1106 on AlpacaEval-2 and a win rate of 46.2\% against GPT-4-0314 on Arena-Hard. Our code is available at https://github.com/SLIT-AI/WRPO.

3277. BBCaL: Black-box Backdoor Detection under the Causality Lens

链接：https://iclr.cc/virtual/2025/poster/31453 abstract：

3278. Reward Dimension Reduction for Scalable Multi-Objective Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28095 abstract： In this paper, we introduce a simple yet effective reward dimension reduction method to tackle the scalability challenges of multi-objective reinforcement learning algorithms. While most existing approaches focus on optimizing two to four objectives, their abilities to scale to environments with more objectives remain uncertain. Our method uses a dimension reduction approach to enhance learning efficiency and policy performance in multi-objective settings. While most traditional dimension reduction methods are designed for static datasets, our approach is tailored for online learning and preserves Pareto-optimality after transformation. We propose a new training and evaluation framework for reward dimension reduction in multi-objective reinforcement learning and demonstrate the superiority of our method in environments including one with sixteen objectives, significantly outperforming existing online dimension reduction methods.

3279. Mind Control through Causal Inference: Predicting Clean Images from Poisoned Data

链接：https://iclr.cc/virtual/2025/poster/28743 abstract： Anti-backdoor learning, aiming to train clean models directly from poisoned datasets, serves as an important defense method for backdoor attack. However, existing methods usually fail to recover backdoored samples to their original, correct labels and suffer from poor generalization to large pre-trained models due to its non end-to end training, making them unsuitable for protecting the increasingly prevalent large pre-trained models. To bridge the gap, we first revisit the anti-backdoor learning problem from a causal perspective. Our theoretical causal analysis reveals that incorporating \emph{\textbf{both}} images and the associated attack indicators preserves the model's integrity. Building on the theoretical analysis, we introduce an end-to-end method, Mind Control through Causal Inference (MCCI), to train clean models directly from poisoned datasets. This approach leverages both the image and the attack indicator to train the model. Based on this training paradigm, the model's perception of whether an input is clean or backdoored can be controlled. Typically, by introducing fake non-attack indicators, the model perceives all inputs as clean and makes correct predictions, even for poisoned samples. Extensive experiments demonstrate that our method achieves state-of-the-art performance, efficiently recovering the original correct predictions for poisoned samples and enhancing accuracy on clean samples.

3280. Equivariant Masked Position Prediction for Efficient Molecular Representation

链接：https://iclr.cc/virtual/2025/poster/29851 abstract： Graph neural networks (GNNs) have shown considerable promise in computational chemistry. However, the limited availability of molecular data raises concerns regarding GNNs' ability to effectively capture the fundamental principles of physics and chemistry, which constrains their generalization capabilities. To address this challenge, we introduce a novel self-supervised approach termed Equivariant Masked Position Prediction (EMPP), grounded in intramolecular potential and force theory. Unlike conventional attribute masking techniques, EMPP formulates a nuanced position prediction task that is more well-defined and enhances the learning of quantum mechanical features. EMPP also bypasses the approximation of the Gaussian mixture distribution commonly used in denoising methods, allowing for more accurate acquisition of physical properties. Experimental results indicate that EMPP significantly enhances performance of advanced molecular architectures, surpassing state-of-the-art self-supervised approaches. Our code is released in https://github.com/ajy112/EMPP.

3281. Tree of Attributes Prompt Learning for Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/27861 abstract： Prompt learning has proven effective in adapting vision language models for downstream tasks. However, existing methods usually append learnable prompt tokens solely with the category names to obtain textual features, which fails to fully leverage the rich context indicated in the category name. To address this issue, we propose the Tree of Attributes Prompt learning (TAP), which first instructs LLMs to generate a tree of attributes with a “concept - attribute - description” structure for each category, and then learn the hierarchy with vision and text prompt tokens. Unlike existing methods that merely augment category names with a set of unstructured descriptions, our approach essentially distills structured knowledge graphs associated with class names from LLMs. Furthermore, our approach introduces text and vision prompts designed to explicitly learn the corresponding visual attributes, effectively serving as domain experts. Additionally, the general and diverse descriptions generated based on the class names may be wrong or absent in the specific given images. To address this misalignment, we further introduce a vision-conditional pooling module to extract instance-specific text features. Extensive experimental results demonstrate that our approach outperforms state-of-the-art methods on the zero-shot base-to-novel generalization, cross-dataset transfer, as well as few-shot classification across 11 diverse datasets. Code is available at https://github.com/HHenryD/TAP.

3282. Decision Information Meets Large Language Models: The Future of Explainable Operations Research

链接：https://iclr.cc/virtual/2025/poster/29386 abstract： Operations Research (OR) is vital for decision-making in many industries. While recent OR methods have seen significant improvements in automation and efficiency through integrating Large Language Models (LLMs), they still struggle to produce meaningful explanations. This lack of clarity raises concerns about transparency and trustworthiness in OR applications. To address these challenges, we propose a comprehensive framework, Explainable Operations Research (EOR), emphasizing actionable and understandable explanations accompanying optimization. The core of EOR is the concept of Decision Information, which emerges from what-if analysis and focuses on evaluating the impact of complex constraints (or parameters) changes on decision-making. Specifically, we utilize bipartite graphs to quantify the changes in the OR model and adopt LLMs to improve the explanation capabilities. Additionally, we introduce the first industrial benchmark to rigorously evaluate the effectiveness of explanations and analyses in OR, establishing a new standard for transparency and clarity in the field.

3283. Streaming Video Question-Answering with In-context Video KV-Cache Retrieval

链接：https://iclr.cc/virtual/2025/poster/30750 abstract： We propose ReKV, a novel training-free approach that enables efficient streaming video question-answering (StreamingVQA), by seamlessly integrating with existing Video Large Language Models (Video-LLMs). Traditional VideoQA systems struggle with long videos, as they must process entire videos before responding to queries, and repeat this process for each new question. In contrast, our approach analyzes long videos in a streaming manner, allowing for prompt responses as soon as user queries are received. Building on a common Video-LLM, we first incorporate a sliding-window attention mechanism, ensuring that input frames attend to a limited number of preceding frames, thereby reducing computational overhead. To prevent information loss, we store processed video key-value caches (KV-Caches) in RAM and disk, reloading them into GPU memory as needed. Additionally, we introduce a retrieval method that leverages an external retriever or the parameters within Video-LLMs to retrieve only query-relevant KV-Caches, ensuring both efficiency and accuracy in question answering. ReKV enables the separation of video analyzing and question-answering across different processes and GPUs, significantly enhancing the efficiency of StreamingVQA. Through comprehensive experimentation, we validate the efficacy and practicality of our approach, which significantly boosts efficiency and enhances applicability over existing VideoQA models.

3284. Sitcom-Crafter: A Plot-Driven Human Motion Generation System in 3D Scenes

链接：https://iclr.cc/virtual/2025/poster/30316 abstract： Recent advancements in human motion synthesis have focused on specific types of motions, such as human-scene interaction, locomotion or human-human interaction, however, there is a lack of a unified system capable of generating a diverse combination of motion types. In response, we introduce Sitcom-Crafter, a comprehensive and extendable system for human motion generation in 3D space, which can be guided by extensive plot contexts to enhance workflow efficiency for anime and game designers. The system is comprised of eight modules, three of which are dedicated to motion generation, while the remaining five are augmentation modules that ensure consistent fusion of motion sequences and system functionality. Central to the generation modules is our novel 3D scene-aware human-human interaction module, which addresses collision issues by synthesizing implicit 3D Signed Distance Function (SDF) points around motion spaces, thereby minimizing human-scene collisions without additional data collection costs. Complementing this, our locomotion and human-scene interaction modules leverage existing methods to enrich the system's motion generation capabilities. Augmentation modules encompass plot comprehension for command generation, motion synchronization for seamless integration of different motion types, hand pose retrieval to enhance motion realism, motion collision revision to prevent human collisions, and 3D retargeting to ensure visual fidelity. Experimental evaluations validate the system's ability to generate high-quality, diverse, and physically realistic motions, underscoring its potential for advancing creative workflows. Code and demonstration videos can be found in the supplementary files.

3285. Explain Yourself, Briefly! Self-Explaining Neural Networks with Concise Sufficient Reasons

链接：https://iclr.cc/virtual/2025/poster/30745 abstract： Minimal sufficient reasons represent a prevalent form of explanation - the smallest subset of input features which, when held constant at their corresponding values, ensure that the prediction remains unchanged. Previous post-hoc methods attempt to obtain such explanations but face two main limitations: (1) Obtaining these subsets poses a computational challenge, leading most scalable methods to converge towards suboptimal, less meaningful subsets; (2) These methods heavily rely on sampling out-of-distribution input assignments, potentially resulting in counterintuitive behaviors. To tackle these limitations, we propose in this work a self-supervised training approach, which we term sufficient subset training (SST). Using SST, we train models to generate concise sufficient reasons for their predictions as an integral part of their output. Our results indicate that our framework produces succinct and faithful subsets substantially more efficiently than competing post-hoc methods while maintaining comparable predictive performance.

3286. GenARM: Reward Guided Generation with Autoregressive Reward Model for Test-Time Alignment

链接：https://iclr.cc/virtual/2025/poster/30136 abstract： Large Language Models (LLMs) exhibit impressive capabilities but require careful alignment with human preferences. Traditional training-time methods finetune LLMs using human preference datasets but incur significant training costs and require repeated training to handle diverse user preferences. Test-time alignment methods address this by using reward models (RMs) to guide frozen LLMs without retraining. However, existing test-time approaches rely on trajectory-level RMs which are designed to evaluate complete responses, making them unsuitable for autoregressive text generation that requires computing next-token rewards from partial responses. To address this, we introduce GenARM, a test-time alignment approach that leverages the Autoregressive Reward Model—a novel reward parametrization designed to predict next-token rewards for efficient and effective autoregressive generation. Theoretically, we demonstrate that this parametrization can provably guide frozen LLMs toward any distribution achievable by traditional RMs within the KL-regularized reinforcement learning framework. Experimental results show that GenARM significantly outperforms prior test-time alignment baselines and matches the performance of training-time methods. Additionally, GenARM enables efficient weak-to-strong guidance, aligning larger LLMs with smaller RMs without the high costs of training larger models. Furthermore, GenARM supports multi-objective alignment, allowing real-time trade-offs between preference dimensions and catering to diverse user preferences without retraining. Our project page is available at: https://genarm.github.io.

3287. Out-of-distribution Generalization for Total Variation based Invariant Risk Minimization

链接：https://iclr.cc/virtual/2025/poster/29073 abstract： Invariant risk minimization is an important general machine learning framework that has recently been interpreted as a total variation model (IRM-TV). However, how to improve out-of-distribution (OOD) generalization in the IRM-TV setting remains unsolved. In this paper, we extend IRM-TV to a Lagrangian multiplier model named OOD-TV-IRM. We find that the autonomous TV penalty hyperparameter is exactly the Lagrangian multiplier. Thus OOD-TV-IRM is essentially a primal-dual optimization model, where the primal optimization minimizes the entire invariant risk and the dual optimization strengthens the TV penalty. The objective is to reach a semi-Nash equilibrium where the balance between the training loss and OOD generalization is maintained. We also develop a convergent primal-dual algorithm that facilitates an adversarial learning scheme. Experimental results show that OOD-TV-IRM outperforms IRM-TV in most situations.

3288. SPD Attack - Prevention of AI Powered Image Editing by Image Immunization

链接：https://iclr.cc/virtual/2025/poster/37631 abstract： Recent advances in image-to-image editing models offer both benefits and risks. While they enhance creativity, accessibility, and applications in fields ranging from medicine to environmental science, they can also enable misuse, such as identity manipulation, copyright infringement, and deepfake creation. This blog explores methods to protect images from such misuse, reproduces findings from relevant research, and extends them across various models and datasets.

3289. CREMA: Generalizable and Efficient Video-Language Reasoning via Multimodal Modular Fusion

链接：https://iclr.cc/virtual/2025/poster/31072 abstract： Despite impressive advancements in recent multimodal reasoning approaches, they are still limited in flexibility and efficiency, as these models typically process only a few fixed modality inputs and require updates to numerous parameters. This paper tackles these critical challenges and proposes CREMA, a generalizable, highly efficient, and modular modality-fusion framework that can incorporate many new modalities to enhance video reasoning. We first augment multiple informative modalities (such as optical flow, 3D point cloud, audio, thermal heatmap, and touch map) from given videos without extra human annotation by leveraging sensors or existing pre-trained models. Next, we introduce a query transformer with multiple parameter-efficient modules associated with each accessible modality. It projects diverse modality features to the LLM token embedding space, allowing the model to integrate different data types for response generation. Furthermore, we propose a novel progressive multimodal fusion design supported by a lightweight fusion module and modality-sequential training strategy. It helps compress information across various assisting modalities, maintaining computational efficiency in the LLM while improving performance. We validate our method on seven video-language reasoning tasks assisted by diverse modalities, including conventional VideoQA and Video-Audio/3D/Touch/Thermal QA, and achieve better/equivalent performance against strong multimodal LLMs, including OneLLM, BLIP-2, and SeViLA while reducing over 90% trainable parameters. We provide extensive analyses of CREMA, including the impact of each modality on reasoning domains, the design of the fusion module, and example visualizations.

3290. Loopy: Taming Audio-Driven Portrait Avatar with Long-Term Motion Dependency

链接：https://iclr.cc/virtual/2025/poster/32049 abstract： With the introduction of video diffusion model, audio-conditioned human video generation has recently achieved significant breakthroughs in both the naturalness of motion and the synthesis of portrait details. Due to the limited control of audio signals in driving human motion, existing methods often add auxiliary spatial signals such as movement regions to stabilize movements, which compromise the naturalness and freedom of motion. To address this issue, we propose an end-to-end audio-only conditioned video diffusion model named Loopy. Specifically, we designed two key modules: an inter- and intra-clip temporal module and an audio-to-latents module. These enable the model to better utilize long-term motion dependencies and establish a stronger audio-portrait movement correlation. Consequently, the model can generate more natural and stable portrait videos with subtle facial expressions, without the need for manually setting movement constraints. Extensive experiments show that Loopy outperforms recent audio-driven portrait diffusion models, delivering more lifelike and high-quality results across various scenarios. Video samples are available at https://loopyavataranony.github.io/

3291. Towards Bridging Generalization and Expressivity of Graph Neural Networks

链接：https://iclr.cc/virtual/2025/poster/30574 abstract： Expressivity and generalization are two critical aspects of graph neural networks (GNNs). While significant progress has been made in studying the expressivity of GNNs, much less is known about their generalization capabilities, particularly when dealing with the inherent complexity of graph-structured data.In this work, we address the intricate relationship between expressivity and generalization in GNNs. Theoretical studies conjecture a trade-off between the two: highly expressive models risk overfitting, while those focused on generalization may sacrifice expressivity. However, empirical evidence often contradicts this assumption, with expressive GNNs frequently demonstrating strong generalization. We explore this contradiction by introducing a novel framework that connects GNN generalization to the variance in graph structures they can capture. This leads us to propose a k-variance margin-based generalization bound that characterizes the structural properties of graph embeddings in terms of their upper-bounded expressive power. Our analysis does not rely on specific GNN architectures, making it broadly applicable across GNN models. We further uncover a trade-off between intra-class concentration and inter-class separation, both of which are crucial for effective generalization. Through case studies and experiments on real-world datasets, we demonstrate that our theoretical findings align with empirical results, offering a deeper understanding of how expressivity can enhance GNN generalization.

3292. Shared-AE: Automatic Identification of Shared Subspaces in High-dimensional Neural and Behavioral Activity

链接：https://iclr.cc/virtual/2025/poster/27666 abstract： Understanding the relationship between behavior and neural activity is crucial for understanding brain function. An effective method is to learn embeddings for interconnected modalities. For simple behavioral tasks, neural features can be learned based on labels. However, complex behaviors, such as social interactions, require the joint extraction of behavioral and neural characteristics. In this paper, we present an autoencoder (AE) framework, called Shared-AE, which includes a novel regularization term that automatically identifies features shared between neural activity and behavior, while simultaneously capturing the unique private features specific to each modality. We apply Shared-AE to large-scale neural activity recorded across the entire dorsal cortex of the mouse, during two very different behaviors: (i) head-fixed mice performing a self-initiated decision-making task, and (ii) freely-moving social behavior amongst two mice. Our model successfully captures both shared features', shared across neural and behavioral activity, andprivate features', unique to each modality, significantly enhancing our understanding of the alignment between neural activity and complex behaviors. The original code for the entire Shared-AE framework on Pytorch has been made publicly available at: \url{https://github.com/saxenalab-neuro/Shared-AE}.

3293. Medium-Difficulty Samples Constitute Smoothed Decision Boundary for Knowledge Distillation on Pruned Datasets

链接：https://iclr.cc/virtual/2025/poster/29633 abstract： This paper tackles a new problem of dataset pruning for Knowledge Distillation (KD), from a fresh perspective of Decision Boundary (DB) preservation and drifts. Existing dataset pruning methods generally assume that the post-pruning DB formed by the selected samples can be well-captured by future networks that use those samples for training. Therefore, they tend to preserve hard samples since hard samples are closer to the DB and better characterize the nuances in the distribution of the entire dataset. However, in KD, the limited learning capacity from the student network leads to imperfect preservation of the teacher's feature distribution, resulting in the drift of DB in the student space. Specifically, hard samples worsen such drifts as they are difficult for the student to learn, creating a situation where the student's DB can drift deeper into other classes and make incorrect classifications. Motivated by these findings, our method selects medium-difficulty samples for KD-based dataset pruning. We show that these samples constitute a smoothed version of the teacher's DB and are easier for the student to learn, obtaining a general feature distribution preservation for a class of samples and reasonable DB between different classes for the student. In addition, to reduce the distributional shift due to dataset pruning, we leverage the class-wise distributional information of the teacher's outputs to reshape the logits of the preserved samples. Experiments show that the proposed static pruning method can even perform better than the state-of-the-art dynamic pruning method which needs access to the entire dataset. In addition, our method halves the training times of KD and improves the student's accuracy by 0.4% on ImageNet with a 50% keep ratio. When the ratio further increases to 70%, our method achieves higher accuracy over the vanilla KD while reducing the training times by 30%. Code is available at https://github.com/chenyd7/MDSLR.

3294. Value-aligned Behavior Cloning for Offline Reinforcement Learning via Bi-level Optimization

链接：https://iclr.cc/virtual/2025/poster/28908 abstract： Offline reinforcement learning (RL) aims to optimize policies under pre-collected data, without requiring any further interactions with the environment. Derived from imitation learning, Behavior cloning (BC) is extensively utilized in offline RL for its simplicity and effectiveness. Although BC inherently avoids out-of-distribution deviations, it lacks the ability to discern between high and low-quality data, potentially leading to sub-optimal performance when facing with poor-quality data. Current offline RL algorithms attempt to enhance BC by incorporating value estimation, yet often struggle to effectively balance these two critical components, specifically the alignment between the behavior policy and the pre-trained value estimations under in-sample offline data. To address this challenge, we propose the Value-aligned Behavior Cloning via Bi-level Optimization (VACO), a novel bi-level framework that seamlessly integrates an inner loop for weighted supervised behavior cloning (BC) with an outer loop dedicated to value alignment. In this framework, the inner loop employs a meta-scoring network to evaluate and appropriately weight each training sample, while the outer loop maximizes value estimation for alignment with controlled noise to facilitate limited exploration. This bi-level structure allows VACO to identify the optimal weighted BC policy, ultimately maximizing the expected estimated return conditioned on the learned value function. We conduct a comprehensive evaluation of VACO across a variety of continuous control benchmarks in offline RL, where it consistently achieves superior performance compared to existing state-of-the-art methods.

3295. Physics-aligned field reconstruction with diffusion bridge

链接：https://iclr.cc/virtual/2025/poster/30484 abstract： The reconstruction of physical fields from sparse measurements is pivotal in both scientific research and engineering applications. Traditional methods are increasingly supplemented by deep learning models due to their efficacy in extracting features from data. However, except for the low accuracy on complex physical systems, these models often fail to comply with essential physical constraints, such as governing equations and boundary conditions. To overcome this limitation, we introduce a novel data-driven field reconstruction framework, termed the Physics-aligned Schr\"{o}dinger Bridge (PalSB). This framework leverages a diffusion bridge mechanism that is specifically tailored to align with physical constraints. The PalSB approach incorporates a dual-stage training process designed to address both local reconstruction mapping and global physical principles. Additionally, a boundary-aware sampling technique is implemented to ensure adherence to physical boundary conditions. We demonstrate the effectiveness of PalSB through its application to three complex nonlinear systems: cylinder flow from Particle Image Velocimetry experiments, two-dimensional turbulence, and a reaction-diffusion system. The results reveal that PalSB not only achieves higher accuracy but also exhibits enhanced compliance with physical constraints compared to existing methods. This highlights PalSB's capability to generate high-quality representations of intricate physical interactions, showcasing its potential for advancing field reconstruction techniques. The source code can be found at https://github.com/lzy12301/PalSB.

3296. How to Find the Exact Pareto Front for Multi-Objective MDPs?

链接：https://iclr.cc/virtual/2025/poster/29630 abstract： Multi-Objective Markov Decision Processes (MO-MDPs) are receiving increasing attention, as real-world decision-making problems often involve conflicting objectives that cannot be addressed by a single-objective MDP. The Pareto front identifies the set of policies that cannot be dominated, providing a foundation for finding Pareto optimal solutions that can efficiently adapt to various preferences.However, finding the Pareto front is a highly challenging problem. Most existing methods either (i) rely on traversing the *continuous preference space*, which is impractical and results in approximations that are difficult to evaluate against the true Pareto front, or (ii) focus solely on deterministic Pareto optimal policies, from which there are no known techniques to characterize the full Pareto front. Moreover, finding the structure of the Pareto front itself remains unclear even in the context of dynamic programming, where the MDP is fully known in advance.In this work, we address the challenge of efficiently discovering the Pareto front, involving both deterministic and stochastic Pareto optimal policies.By investigating the geometric structure of the Pareto front in MO-MDPs, we uncover a key property: the Pareto front is on the boundary of a convex polytope whose vertices all correspond to deterministic policies, and neighboring vertices of the Pareto front differ by only one state-action pair of the deterministic policy, almost surely.This insight transforms the global comparison across all policies into a localized search among deterministic policies that differ by only one state-action pair, drastically reducing the complexity of searching for the exact Pareto front. We develop an efficient algorithm that identifies the vertices of the Pareto front by solving a single-objective MDP only once and then traversing the edges of the Pareto front, making it more efficient than existing methods. Furthermore, the entire Pareto front can be found in V iterations, where V represents the number of vertices on the Pareto front.Our empirical studies demonstrate the effectiveness of our theoretical strategy in discovering the Pareto front efficiently.

3297. Hybrid Regularization Improves Diffusion-based Inverse Problem Solving

链接：https://iclr.cc/virtual/2025/poster/29007 abstract： Diffusion models, recognized for their effectiveness as generative priors, have become essential tools for addressing a wide range of visual challenges. Recently, there has been a surge of interest in leveraging Denoising processes for Regularization (DR) to solve inverse problems. However, existing methods often face issues such as mode collapse, which results in excessive smoothing and diminished diversity. In this study, we perform a comprehensive analysis to pinpoint the root causes of gradient inaccuracies inherent in DR. Drawing on insights from diffusion model distillation, we propose a novel approach called Consistency Regularization (CR), which provides stabilized gradients without the need for ODE simulations. Building on this, we introduce Hybrid Regularization (HR), a unified framework that combines the strengths of both DR and CR, harnessing their synergistic potential. Our approach proves to be effective across a broad spectrum of inverse problems, encompassing both linear and nonlinear scenarios, as well as various measurement noise statistics. Experimental evaluations on benchmark datasets, including FFHQ and ImageNet, demonstrate that our proposed framework not only achieves highly competitive results compared to state-of-the-art methods but also offers significant reductions in wall-clock time and memory consumption.

3298. Broadening Target Distributions for Accelerated Diffusion Models via a Novel Analysis Approach

链接：https://iclr.cc/virtual/2025/poster/28182 abstract： Accelerated diffusion models hold the potential to significantly enhance the efficiency of standard diffusion processes. Theoretically, these models have been shown to achieve faster convergence rates than the standard $\mathcal O(1/\epsilon^2)$ rate of vanilla diffusion models, where ϵ denotes the target accuracy. However, current theoretical studies have established the acceleration advantage only for restrictive target distribution classes, such as those with smoothness conditions imposed along the entire sampling path or with bounded support. In this work, we significantly broaden the target distribution classes with a new accelerated stochastic DDPM sampler. In particular, we show that it achieves accelerated performance for three broad distribution classes not considered before. Our first class relies on the smoothness condition posed only to the target density q_0, which is far more relaxed than the existing smoothness conditions posed to all q_t along the entire sampling path. Our second class requires only a finite second moment condition, allowing for a much wider class of target distributions than the existing finite-support condition. Our third class is Gaussian mixture, for which our result establishes the first acceleration guarantee. Moreover, among accelerated DDPM type samplers, our results specialized for bounded-support distributions show an improved dependency on the data dimension d. Our analysis introduces a novel technique for establishing performance guarantees via constructing a tilting factor representation of the convergence error and utilizing Tweedie's formula to handle Taylor expansion terms. This new analytical framework may be of independent interest.

3299. Theory on Score-Mismatched Diffusion Models and Zero-Shot Conditional Samplers

链接：https://iclr.cc/virtual/2025/poster/28407 abstract： The denoising diffusion model has recently emerged as a powerful generative technique, capable of transforming noise into meaningful data. While theoretical convergence guarantees for diffusion models are well established when the target distribution aligns with the training distribution, practical scenarios often present mismatches. One common case is in the zero-shot conditional diffusion sampling, where the target conditional distribution is different from the (unconditional) training distribution. These score-mismatched diffusion models remain largely unexplored from a theoretical perspective. In this paper, we present the first performance guarantee with explicit dimensional dependencies for general score-mismatched diffusion samplers, focusing on target distributions with finite second moments. We show that score mismatches result in an asymptotic distributional bias between the target and sampling distributions, proportional to the accumulated mismatch between the target and training distributions. This result can be directly applied to zero-shot conditional samplers for any conditional model, irrespective of measurement noise. Interestingly, the derived convergence upper bound offers useful guidance for designing a novel bias-optimal zero-shot sampler in linear conditional models that minimizes the asymptotic bias. For such bias-optimal samplers, we further establish convergence guarantees with explicit dependencies on dimension and conditioning, applied to several interesting target distributions, including those with bounded support and Gaussian mixtures. Our findings are supported by numerical studies.

3300. PEARL: Towards Permutation-Resilient LLMs

链接：https://iclr.cc/virtual/2025/poster/28005 abstract： The in-context learning (ICL) capability of large language models (LLMs) enables them to perform challenging tasks using provided demonstrations. However, ICL is highly sensitive to the ordering of demonstrations, leading to instability in predictions. This paper shows that this vulnerability can be exploited to design a natural attack—difficult for model providers to detect—that achieves nearly 80% success rate on LLaMA-3 by simply permuting the demonstrations. Existing mitigation methods primarily rely on post-processing and fail to enhance the model's inherent robustness to input permutations, raising concerns about safety and reliability of LLMs. To address this issue, we propose Permutation-resilient learning (PEARL), a novel framework based on distributionally robust optimization (DRO), which optimizes model performance against the worst-case input permutation. Specifically, PEARL consists of a permutation-proposal network (P-Net) and the LLM. The P-Net generates the most challenging permutations by treating it as an optimal transport problem, which is solved using an entropy-constrained Sinkhorn algorithm. Through minimax optimization, the P-Net and the LLM iteratively optimize against each other, progressively improving the LLM's robustness. Experiments on synthetic pre-training and real-world instruction tuning tasks demonstrate that PEARL effectively mitigates permutation attacks and enhances performance. Notably, despite being trained on fewer shots and shorter contexts, PEARL achieves performance gains of up to 40% when scaled to many-shot and long-context scenarios, highlighting its efficiency and generalization capabilities.

3301. SymmetricDiffusers: Learning Discrete Diffusion on Finite Symmetric Groups

链接：https://iclr.cc/virtual/2025/poster/30408 abstract： The group of permutations S_n, also known as the finite symmetric groups, are essential in fields such as combinatorics, physics, and chemistry. However, learning a probability distribution over S_n poses significant challenges due to its intractable size and discrete nature. In this paper, we introduce *SymmetricDiffusers*, a novel discrete diffusion model that simplifies the task of learning a complicated distribution over S_n by decomposing it into learning simpler transitions of the reverse diffusion using deep neural networks. We identify the riffle shuffle as an effective forward transition and provide empirical guidelines for selecting the diffusion length based on the theory of random walks on finite groups. Additionally, we propose a generalized Plackett-Luce (PL) distribution for the reverse transition, which is provably more expressive than the PL distribution. We further introduce a theoretically grounded "denoising schedule" to improve sampling and learning efficiency. Extensive experiments show that our model achieves state-of-the-art or comparable performance on solving tasks including sorting 4-digit MNIST images, jigsaw puzzles, and traveling salesman problems. Our code is released at .

3302. CaPo: Cooperative Plan Optimization for Efficient Embodied Multi-Agent Cooperation

链接：https://iclr.cc/virtual/2025/poster/30059 abstract： In this work, we address the cooperation problem among large language model (LLM) based embodied agents, where agents must cooperate to achieve a common goal. Previous methods often execute actions extemporaneously and incoherently, without long-term strategic and cooperative planning, leading to redundant steps, failures, and even serious repercussions in complex tasks like search-and-rescue missions where discussion and cooperative plan are crucial. To solve this issue, we propose Cooperative Plan Optimization (CaPo) to enhance the cooperation efficiency of LLM-based embodied agents. Inspired by human cooperation schemes, CaPo improves cooperation efficiency with two phases: 1) meta plan generation, and 2) progress-adaptive meta plan and execution. In the first phase, all agents analyze the task, discuss, and cooperatively create a meta-plan that decomposes the task into subtasks with detailed steps, ensuring a long-term strategic and coherent plan for efficient coordination. In the second phase, agents execute tasks according to the meta-plan and dynamically adjust it based on their latest progress (e.g., discovering a target object) through multi-turn discussions. This progress-based adaptation eliminates redundant actions, improving the overall cooperation efficiency of agents. Experimental results on the ThreeDworld Multi-Agent Transport and Communicative Watch-And-Help tasks demonstrate CaPo's much higher task completion rate and efficiency compared with state-of-the-arts. The code is released at https://github.com/jliu4ai/CaPo.

3303. OpenMathInstruct-2: Accelerating AI for Math with Massive Open-

Source Instruction Data

链接：https://iclr.cc/virtual/2025/poster/28464 abstract： Mathematical reasoning continues to be a critical challenge in large language model (LLM) development with significant interest. However, most of the cutting-edge progress in mathematical reasoning with LLMs has become closed-source due to lack of access to training data. This lack of data access limits researchers from understanding the impact of different choices for synthesizing and utilizing the data. With the goal of creating a high-quality finetuning (SFT) dataset for math reasoning, we conduct careful ablation experiments on data synthesis using the recently released Llama3.1 family of models. Our experiments show that: (a) solution format matters, with excessively verbose solutions proving detrimental to SFT performance, (b) data generated by a strong teacher outperforms on-policy data generated by a weak student model, (c) SFT is robust to low-quality solutions, allowing for imprecise data filtering, and (d) question diversity is crucial for achieving data scaling gains. Based on these insights, we create the OpenMathInstruct-2 dataset which consists of 14M question-solution pairs (≈ 600K unique questions), making it nearly eight times larger than the previous largest open-source math reasoning dataset. Finetuning the Llama-3.1-8B-Base using OpenMathInstruct-2 outperforms Llama3.1-8B-Instruct on MATH by an absolute 15.9% (51.9% → 67.8%). Finally, to accelerate the open-source efforts, we release the code, the finetuned models, and the OpenMathInstruct-2 dataset under a commercially permissive license.

3304. HGM^3: Hierarchical Generative Masked Motion Modeling with Hard Token Mining

链接：https://iclr.cc/virtual/2025/poster/30181 abstract： Text-to-motion generation has significant potential in a wide range of applications including animation, robotics, and AR/VR. While recent works on masked motion models are promising, the task remains challenging due to the inherent ambiguity in text and the complexity of human motion dynamics. To overcome the issues, we propose a novel text-to-motion generation framework that integrates two key components: Hard Token Mining (HTM) and a Hierarchical Generative Masked Motion Model (HGM^3). Our HTM identifies and masks challenging regions in motion sequences and directs the model to focus on hard-to-learn components for efficacy. Concurrently, the hierarchical model uses a semantic graph to represent sentences at different granularity, allowing the model to learn contextually feasible motions. By leveraging a shared-weight masked motion model, it reconstructs the same sequence under different conditioning levels and facilitates comprehensive learning of complex motion patterns. During inference, the model progressively generates motions by incrementally building up coarse-to-fine details. Extensive experiments on benchmark datasets, including HumanML3D and KIT-ML, demonstrate that our method outperforms existing methods in both qualitative and quantitative measures for generating context-aware motions.

3305. G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model

链接：https://iclr.cc/virtual/2025/poster/28274 abstract： Large language models (LLMs) have shown remarkable proficiency in human-level reasoning and generation capabilities, which encourages extensive research on their application in mathematical problem solving. However, current work has been largely focused on text-based mathematical problems, with limited investigation in problems involving multi-modal geometric information. Addressing this gap, we aim to enable LLMs to solve geometric problems by understanding image input. We first identify the limitations of current Multimodal Large Language Models (MLLMs) in this area: they struggle to accurately comprehend basic geometric elements and their relationships. To address these challenges, we leverage the inherent attribute of logical structure compactness in geometric figures, utilizing text-only Large Language Models (LLMs) to curate a comprehensive multimodal geometry dataset. This dataset, named Geo170k, contains more than 170K geometric image-caption and question-answer pairs. Utilizing the Geo170k dataset, we introduce G-LLaVA, a model that demonstrates exceptional performance in solving geometric problems. It significantly outperforms GPT4-V on the geometry task of MathVista benchmark with only 7B parameters.

3306. MTU-Bench: A Multi-granularity Tool-Use Benchmark for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30871 abstract： Large Language Models (LLMs) have displayed massive improvements in reason- ing and decision-making skills and can hold natural conversations with users. Recently, many tool-use benchmark datasets have been proposed. However, existing datasets have the following limitations: (1). Insufficient evaluation scenarios (e.g., only cover limited tool-use scenes). (2). Extensive evaluation costs (e.g., GPT API costs). To address these limitations, in this work, we propose a multi-granularity tool-use benchmark for large language models called MTU-Bench. For the "multi-granularity" property, our MTU-Bench covers five tool usage scenes (i.e., single-turn and single-tool, single-turn and multiple-tool, multiple-turn and single-tool, multiple-turn and multiple-tool, and out-of-distribution tasks). Besides, all evaluation metrics of our MTU-Bench are based on the prediction results and the ground truth without using any GPT or human evaluation metrics. Moreover, our MTU-Bench is collected by transforming existing high-quality datasets to simulate real-world tool usage scenarios, and we also propose an instruction dataset called MTU-Instruct data to enhance the tool-use abilities of existing LLMs. Comprehensive experimental results demonstrate the effectiveness of our MTU-Bench.

3307. Logic-Logit: A Logic-Based Approach to Choice Modeling

链接：https://iclr.cc/virtual/2025/poster/27923 abstract： In this study, we propose a novel rule-based interpretable choice model, {\bf Logic-Logit}, designed to effectively learn and explain human choices. Choice models have been widely applied across various domains—such as commercial demand forecasting, recommendation systems, and consumer behavior analysis—typically categorized as parametric, nonparametric, or deep network-based. While recent innovations have favored neural network approaches for their computational power, these flexible models often involve large parameter sets and lack interpretability, limiting their effectiveness in contexts where transparency is essential.Previous empirical evidence shows that individuals usually use {\it heuristic decision rules} to form their consideration sets, from which they then choose. These rules are often represented as {\it disjunctions of conjunctions} (i.e., OR-of-ANDs). These rules-driven, {\it consider-then-choose} decision processes enable people to quickly screen numerous alternatives while reducing cognitive and search costs. Motivated by this insight, our approach leverages logic rules to elucidate human choices, providing a fresh perspective on preference modeling. We introduce a unique combination of column generation techniques and the Frank-Wolfe algorithm to facilitate efficient rule extraction for preference modeling—a process recognized as NP-hard. Our empirical evaluation, conducted on both synthetic datasets and real-world data from commercial and healthcare domains, demonstrates that Logic-Logit significantly outperforms baseline models in terms of interpretability and accuracy.

3308. Learning Evolving Tools for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/27814 abstract： Tool learning enables large language models (LLMs) to interact with external tools and APIs, greatly expanding the application scope of LLMs. However, due to the dynamic nature of external environments, these tools and APIs may become outdated over time, preventing LLMs from correctly invoking tools. Existing research primarily focuses on static environments and overlooks this issue, limiting the adaptability of LLMs in real-world applications. In this paper, we propose ToolEVO, a novel framework designed to enhance the adaptive and reflective capabilities of LLMs against tool variability. By leveraging Monte Carlo Tree Search, ToolEVO facilitates active exploration and interaction of LLMs within dynamic environments, allowing for autonomous self-reflection and self-updating of tool usage based on environmental feedback. Additionally, we introduce ToolQA-D, a benchmark specifically designed to evaluate the impact of tool variability. Extensive experiments demonstrate the effectiveness and stability of our approach, highlighting the importance of adaptability to tool variability for effective tool learning.

3309. Enhancing Compositional Text-to-Image Generation with Reliable Random Seeds

链接：https://iclr.cc/virtual/2025/poster/30962 abstract： Text-to-image diffusion models have demonstrated remarkable capability in generating realistic images from arbitrary text prompts. However, they often produce inconsistent results for compositional prompts such as "two dogs" or "a penguin on the right of a bowl". Understanding these inconsistencies is crucial for reliable image generation. In this paper, we highlight the significant role of initial noise in these inconsistencies, where certain noise patterns are more reliable for compositional prompts than others. Our analyses reveal that different initial random seeds tend to guide the model to place objects in distinct image areas, potentially adhering to specific patterns of camera angles and image composition associated with the seed. To improve the model's compositional ability, we propose a method for mining these reliable cases, resulting in a curated training set of generated images without requiring any manual annotation. By fine-tuning text-to-image models on these generated images, we significantly enhance their compositional capabilities. For numerical composition, we observe relative increases of 29.3\% and 19.5\% for Stable Diffusion and PixArt-α, respectively. Spatial composition sees even larger gains, with 60.7\% for Stable Diffusion and 21.1\% for PixArt-α.

3310. WorkflowLLM: Enhancing Workflow Orchestration Capability of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/31085 abstract： Recent advancements in large language models (LLMs) have driven a revolutionary paradigm shift in process automation from Robotic Process Automation to Agentic Process Automation by automating the workflow orchestration procedure based on LLMs. However, existing LLMs (even the advanced OpenAI GPT-4o) are confined to achieving satisfactory capability in workflow orchestration. To address this limitation, we present WorkflowLLM, a data-centric framework elaborately designed to enhance the capability of LLMs in workflow orchestration. It first constructs a large-scale fine-tuning dataset WorkflowBench with 106, 763 samples, covering 1, 503 APIs from 83 applications across 28 categories. Specifically, the construction process can be divided into three phases: (1) Data Collection: we collect real-world workflow data from Apple Shortcuts and RoutineHub, transcribing them into Python-style code. We further equip them with generated hierarchical thought via GPT-4o-mini. (2) Query Expansion: we prompt GPT-4o-mini to generate more task queries to enrich the diversity and complexity of workflows. (3) Workflow Generation: we leverage an annotator model trained on collected data to generate workflows for synthesized queries. Finally, we merge the synthetic samples that pass quality confirmation with the collected samples to obtain the WorkflowBench. Based on WorkflowBench, we fine-tune Llama-3.1-8B to obtain WorkflowLlama. Our experiments show that WorkflowLlama demonstrates a strong capacity to orchestrate complex workflows, while also achieving notable generalization performance on previously unseen APIs. Additionally, WorkflowBench exhibits robust zero-shot generalization capabilities on an out-of-distribution task planning dataset, T-Eval. Our data and code are available at https://github.com/OpenBMB/WorkflowLLM.

3311. Proactive Agent: Shifting LLM Agents from Reactive Responses to

Active Assistance

链接：https://iclr.cc/virtual/2025/poster/28128 abstract： Agents powered by large language models have shown remarkable abilities in solving complex tasks. However, most agent systems remain reactive, limiting their effectiveness in scenarios requiring foresight and autonomous decision-making. In this paper, we tackle the challenge of developing proactive agents capable of anticipating and initiating tasks without explicit human instructions. We propose a novel data-driven approach for this problem. Firstly, we collect real-world human activities to generate proactive task predictions. These predictions are then labeled by human annotators as either accepted or rejected. The labeled data is used to train a reward model that simulates human judgment and serves as an automatic evaluator of the proactiveness of LLM agents. Building on this, we develop a comprehensive data generation pipeline to create a diverse dataset, ProactiveBench, containing 6,790 events. Finally, we demonstrate that fine-tuning models with the proposed ProactiveBench can significantly elicit the proactiveness of LLM agents. Experimental results show that our fine-tuned model achieves an F1-Score of 66.47% in proactively offering assistance, outperforming all open-source and close-source models. These results highlight the potential of our method in creating more proactive and effective agent systems, paving the way for future advancements in human-agent collaboration.

3312. Robust Representation Consistency Model via Contrastive Denoising

链接：https://iclr.cc/virtual/2025/poster/29144 abstract： Robustness is essential for deep neural networks, especially in security-sensitive applications. To this end, randomized smoothing provides theoretical guarantees for certifying robustness against adversarial perturbations. Recently, diffusion models have been successfully employed for randomized smoothing to purify noise-perturbed samples before making predictions with a standard classifier. While these methods excel at small perturbation radii, they struggle with larger perturbations and incur a significant computational overhead during inference compared to classical methods. To address this, we reformulate the generative modeling task along the diffusion trajectories in pixel space as a discriminative task in the latent space. Specifically, we use instance discrimination to achieve consistent representations along the trajectories by aligning temporally adjacent points. After fine-tuning based on the learned representations, our model enables implicit denoising-then-classification via a single prediction, substantially reducing inference costs. We conduct extensive experiments on various datasets and achieve state-of-the-art performance with minimal computation budget during inference. For example, our method outperforms the certified accuracy of diffusion-based methods on ImageNet across all perturbation radii by 5.3\% on average, with up to 11.6\% at larger radii, while reducing inference costs by 85x on average.

3313. Frequency-Guided Masking for Enhanced Vision Self-Supervised Learning

链接：https://iclr.cc/virtual/2025/poster/29401 abstract： We present a novel frequency-based Self-Supervised Learning (SSL) approach that significantly enhances its efficacy for pre-training. Prior work in this direction masks out pre-defined frequencies in the input image and employs a reconstruction loss to pre-train the model. While achieving promising results, such an implementation has two fundamental limitations as identified in our paper. First, using pre-defined frequencies overlooks the variability of image frequency responses. Second, pre-trained with frequency-filtered images, the resulting model needs relatively more data to adapt to naturally looking images during fine-tuning. To address these drawbacks, we propose FOurier transform compression with seLf-Knowledge distillation (FOLK), integrating two dedicated ideas. First, inspired by image compression, we adaptively select the masked-out frequencies based on image frequency responses, creating more suitable SSL tasks for pre-training. Second, we employ a two-branch framework empowered by knowledge distillation, enabling the model to take both the filtered and original images as input, largely reducing the burden of downstream tasks. Our experimental results demonstrate the effectiveness of FOLK in achieving competitive performance to many state-of-the-art SSL methods across various downstream tasks, including image classification, few-shot learning, and semantic segmentation.

3314. Tracing Representation Progression: Analyzing and Enhancing Layer-Wise Similarity

链接：https://iclr.cc/virtual/2025/poster/27907 abstract： Analyzing the similarity of internal representations within and across different models has been an important technique for understanding the behavior of deep neural networks. Most existing methods for analyzing the similarity between representations of high dimensions, such as those based on Centered Kernel Alignment (CKA), rely on statistical properties of the representations for a set of data points. In this paper, we focus on transformer models and study the similarity of representations between the hidden layers of individual transformers. In this context, we show that a simple sample-wise cosine similarity metric is capable of capturing the similarity and aligns with the complicated CKA. Our experimental results on common transformers reveal that representations across layers are positively correlated, with similarity increasing when layers get closer. We provide a theoretical justification for this phenomenon under the geodesic curve assumption for the learned transformer, a property that may approximately hold for residual networks. We then show that an increase in representation similarity implies an increase in predicted probability when directly applying the last-layer classifier to any hidden layer representation. This offers a justification for {\it saturation events}, where the model's top prediction remains unchanged across subsequent layers, indicating that the shallow layer has already learned the necessary knowledge. We then propose an aligned training method to improve the effectiveness of shallow layer by enhancing the similarity between internal representations, with trained models that enjoy the following properties: (1) more early saturation events, (2) layer-wise accuracies monotonically increase and reveal the minimal depth needed for the given task, (3) when served as multi-exit models, they achieve on-par performance with standard multi-exit architectures which consist of additional classifiers designed for early exiting in shallow layers. To our knowledge, our work is the first to show that one common classifier is sufficient for multi-exit models. We conduct experiments on both vision and NLP tasks to demonstrate the performance of the proposed aligned training.

3315. Robustness Reprogramming for Representation Learning

链接：https://iclr.cc/virtual/2025/poster/29574 abstract： This work tackles an intriguing and fundamental open challenge in representation learning: Given a well-trained deep learning model, can it be reprogrammed to enhance its robustness against adversarial or noisy input perturbations without altering its parameters?To explore this, we revisit the core feature transformation mechanism in representation learning and propose a novel non-linear robust pattern matching technique as a robust alternative. Furthermore, we introduce three model reprogramming paradigms to offer flexible control of robustness under different efficiency requirements. Comprehensive experiments and ablation studies across diverse learning models ranging from basic linear model and MLPs to shallow and modern deep ConvNets demonstrate the effectiveness of our approaches.This work not only opens a promising and orthogonal direction for improving adversarial defenses in deep learning beyond existing methods but also provides new insights into designing more resilient AI systems with robust statistics. Our implementation is available at https://github.com/chris-hzc/Robustness-Reprogramming.

3316. On the Linear Speedup of Personalized Federated Reinforcement Learning with Shared Representations

链接：https://iclr.cc/virtual/2025/poster/30560 abstract： Federated reinforcement learning (FedRL) enables multiple agents to collaboratively learn a policy without needing to share the local trajectories collected during agent-environment interactions. However, in practice, the environments faced by different agents are often heterogeneous, but since existing FedRL algorithms learn a single policy across all agents, this may lead to poor performance. In this paper, we introduce a personalized FedRL framework (PFedRL) by taking advantage of possibly shared common structure among agents in heterogeneous environments. Specifically, we develop a class of PFedRL algorithms named PFedRL-Rep that learns (1) a shared feature representation collaboratively among all agents, and (2) an agent-specific weight vector personalized to its local environment. We analyze the convergence of PFedTD-Rep, a particular instance of the framework with temporal difference (TD) learning and linear representations. To the best of our knowledge, we are the first to prove a linear convergence speedup with respect to the number of agents in the PFedRL setting. To achieve this, we show that PFedTD-Rep is an example of federated two-timescale stochastic approximation with Markovian noise. Experimental results demonstrate that PFedTD-Rep, along with an extension to the control setting based on deep Q-networks (DQN), not only improve learning in heterogeneous settings, but also provide better generalization to new environments.

3317. Understanding Constraint Inference in Safety-Critical Inverse Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/30596 abstract： In practical applications, the underlying constraint knowledge is often unknown and difficult to specify. To address this issue, recent advances in Inverse Constrained Reinforcement Learning (ICRL) have focused on inferring these constraints from expert demonstrations. However, the ICRL approach typically characterizes constraint learning as a tri-level optimization problem, which is inherently complex due to its interdependent variables and multiple layers of optimization.Considering these challenges, a critical question arises: *Can we implicitly embed constraint signals into reward functions and effectively solve this problem using a classic reward inference algorithm?* The resulting method, known as Inverse Reward Correction (IRC), merits investigation. In this work, we conduct a theoretical analysis comparing the sample complexities of both solvers. Our findings confirm that the IRC solver achieves lower sample complexity than its ICRL counterpart.Nevertheless, this reduction in complexity comes at the expense of generalizability. Specifically, in the target environment, the reward correction terms may fail to guarantee the safety of the resulting policy, whereas this issue can be effectively mitigated by transferring the constraints via the ICRL solver. Advancing our inquiry, we investigate conditions under which the ICRL solver ensures ϵ-optimality when transferring to new environments. Empirical results across various environments validate our theoretical findings, underscoring the nuanced trade-offs between complexity reduction and generalizability in safety-critical applications.

3318. Alchemy: Amplifying Theorem-Proving Capability Through Symbolic Mutation

链接：https://iclr.cc/virtual/2025/poster/30827 abstract： Formal proofs are challenging to write even for experienced experts. Recent progress in Neural Theorem Proving (NTP) shows promise in expediting this process. However, the formal corpora available on the Internet are limited compared to the general text, posing a significant data scarcity challenge for NTP. To address this issue, this work proposes Alchemy, a general framework for data synthesis that constructs formal theorems through symbolic mutation. Specifically, for each candidate theorem in Mathlib, we identify all invocable theorems that can be used to rewrite or apply to it. Subsequently, we mutate the candidate theorem by replacing the corresponding term in the statement with its equivalent form or antecedent. As a result, our method increases the number of theorems in Mathlib by an order of magnitude, from 110k to 6M. Furthermore, we perform continual pretraining and supervised finetuning on this augmented corpus for large language models. Experimental results demonstrate the effectiveness of our approach, achieving a 4.70% absolute performance improvement on Leandojo benchmark. Additionally, our approach achieves a 2.47% absolute performance gain on the out-of-distribution miniF2F benchmark based on the synthetic data. To provide further insights, we conduct a comprehensive analysis of synthetic data composition and the training paradigm, offering valuable guidance for developing a strong theorem prover.

3319. Multi-Accurate CATE is Robust to Unknown Covariate Shifts

链接：https://iclr.cc/virtual/2025/poster/31461 abstract： Estimating heterogeneous treatment effects is important to tailor treatments to those individuals who would most likely benefit. However, conditional average treatment effect predictors may often be trained on one population but possibly deployed on different, possibly unknown populations. We use methodology for learning multi-accurate predictors to post-process CATE T-learners (differenced regressions) to become robust to unknown covariate shifts at the time of deployment. The method works in general for pseudo-outcome regression, such as the DR-learner. We show how this approach can combine (large) confounded observational and (smaller) randomized datasets by learning a confounded predictor from the observational dataset, and auditing for multi-accuracy on the randomized controlled trial. We show improvements in bias and mean squared error in simulations with increasingly larger covariate shift, and on a semi-synthetic case study of a parallel large observational study and smaller randomized controlled experiment. Overall, we establish a connection between methods developed for multi-distribution learning and achieve appealing desiderata (e.g. external validity) in causal inference and machine learning.

3320. OpenPRM: Building Open-domain Process-based Reward Models with Preference Trees

链接：https://iclr.cc/virtual/2025/poster/28887 abstract： Scaling inference-time computation is increasingly seen as the next frontier in scaling laws for large language models. Previous work in mathematics and coding has demonstrated the remarkable potential for inference-time scaling. During such scaling, fine-grained supervision through process-based reward models (PRMs) is essential for enhancement. However, exploration of inference-time scaling and PRMs in open-domain problems remains limited, where lacking exact answers and obtaining process supervision prove challenging. In this paper, we explore the construction of PRMs for open-domain tasks, specifically for instruction-following tasks. Utilizing existing outcome-based reward models (ORMs), we develop sentence-level preference trees based on the prefix similarity of parallel sampled candidates from datasets like UltraFeedback. This setup allows us to derive weak supervision for processes via back-propagation from outcome-level rewards. Subsequently, we integrate ORMs and PRMs under the same pairwise ranking objectives, resulting in our newly developed reward models, named OpenPRM. This approach significantly enhances the scalability of process-level supervision in open domains at minimal cost. We assess the performance of OpenPRM across various reward benchmarks, demonstrating its competitive edge over traditional ORMs in open domains and PRMs in specialized domains. Additionally, we investigate the scalability of inference-time computation for open-domain instructions. Our results highlight the limitations of ORMs’ scalability, while OpenPRM shows superior performance in scaled settings. Despite these advances, achieving automatic fine-grained supervision for open-domain inference-time scaling remains a substantial challenge. We hope these findings will spur further development of process supervision reward models in open-domain scenarios.

3321. Advancing LLM Reasoning Generalists with Preference Trees

链接：https://iclr.cc/virtual/2025/poster/31127 abstract： We introduce EURUS, a suite of large language models (LLMs) optimized for reasoning. Finetuned from Mistral-7B, Llama-3-8B, and Mixtral-8x22B, EURUS models achieve state-of-the-art results among open-source models on a diverse set of benchmarks covering mathematics, code generation, and logical reasoning problems. Notably, EURUX-8X22B outperforms GPT-3.5 Turbo in reasoning through a comprehensive benchmarking across 12 test sets covering five tasks. The strong performance of EURUS can be primarily attributed to ULTRAINTERACT, our newly-curated large-scale, high-quality training data dataset specifically designed for complex reasoning tasks. ULTRAINTERACT can be used in both supervised fine-tuning, preference learning, and reward modeling. It pairs each instruction with a preference tree consisting of (1) reasoning chains with diverse planning strategies in a unified format, (2) multi-turn interaction trajectories with the environment and the critique, and (3) pairwise positive and negative responses to facilitate preference learning. ULTRAINTERACT allows us to conduct an in-depth exploration of preference learning for reasoning tasks. Our investigation reveals that some well-established preference learning algorithms may be less suitable for reasoning tasks compared to their effectiveness in general conversations. The hypothesis is that in reasoning tasks, the space of correct answers is much smaller than that of incorrect ones, so it is necessary to explicitly increase the reward of chosen data. Therefore, in addition to increasing the reward margin as many preference learning algorithms do, the absolute values of positive responses’ rewards should be positive and may serve as a proxy for performance. Inspired by this, we derive a novel reward modeling objective and empirically that it leads to a stable reward modeling curve and better performance. Together with ULTRAINTERACT, we obtain a strong reward model.

3322. NeuralPlane: Structured 3D Reconstruction in Planar Primitives with Neural Fields

链接：https://iclr.cc/virtual/2025/poster/30944 abstract： 3D maps assembled from planar primitives are compact and expressive in representing man-made environments. In this paper, we present NeuralPlane, a novel approach that explores neural fields for multi-view 3D plane reconstruction. Our method is centered upon the core idea of distilling geometric and semantic cues from inconsistent 2D plane observations into a unified 3D neural representation, which unlocks the full leverage of plane attributes. It is accomplished through several key designs, including: 1) a monocular module that generates geometrically smooth and semantically meaningful segments known as 2D plane observations, 2) a plane-guided training procedure that implicitly learns accurate 3D geometry from the multi-view plane observations, and 3) a self-supervised feature field termed Neural Coplanarity Field that enables the modeling of scene semantics alongside the geometry. Without relying on prior plane annotations, our method achieves high-fidelity reconstruction comprising planar primitives that are not only crisp but also well-aligned with the semantic content. Comprehensive experiments on ScanNetv2 and ScanNet++ demonstrate the superiority of our method in both geometry and semantics.

3323. Self-play with Execution Feedback: Improving Instruction-following Capabilities of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29049 abstract： One core capability of large language models~(LLMs) is to follow natural language instructions. However, the issue of automatically constructing high-quality training data to enhance the complex instruction-following abilities of LLMs without manual annotation remains unresolved. In this paper, we introduce AutoIF, the first scalable and reliable method for automatically generating instruction-following training data. AutoIF transforms the validation of instruction-following data quality into code verification, requiring LLMs to generate instructions, the corresponding code to verify the correctness of the instruction responses, and unit test samples to cross-validate the code's correctness. Then, execution feedback-based rejection sampling can generate data for Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) training. AutoIF achieves significant improvements across three training algorithms, SFT, Offline DPO, and Online DPO, when applied to the advanced open-source LLMs, Qwen2 and LLaMA3, in self-alignment and strong-to-weak distillation settings. Using two widely-used and three challenging general instruction-following benchmarks, we demonstrate that AutoIF significantly improves LLM performance across a wide range of natural instruction constraints. Notably, AutoIF is the first to surpass 90\% accuracy in IFEval’s loose instruction accuracy, without compromising general, math and coding capabilities. Further analysis of quality, scaling, combination, and data efficiency highlights AutoIF's strong generalization and alignment potential. Our code are available at https://github.com/QwenLM/AutoIF

3324. Accelerating Inference of Retrieval-Augmented Generation via Sparse Context Selection

链接：https://iclr.cc/virtual/2025/poster/30237 abstract： Large language models (LLMs) augmented with retrieval exhibit robust performance and extensive versatility by incorporating external contexts. However, the input length grows linearly in the number of retrieved documents, causing a dramatic increase in latency. In this paper, we propose a novel paradigm named Sparse RAG, which seeks to cut computation costs through sparsity. Specifically, Sparse RAG encodes retrieved documents in parallel, which eliminates latency introduced by long-range attention of retrieved documents. Then, LLMs selectively decode the output by only attending to highly relevant caches auto-regressively, which are chosen via prompting LLMs with special control tokens. It is notable that Sparse RAG combines the assessment of each individual document and the generation of the response into a single process. The designed sparse mechanism in a RAG system can facilitate the reduction of the number of documents loaded during decoding for accelerating the inference of the RAG system. Additionally, filtering out undesirable contexts enhances the model’s focus on relevant context, inherently improving its generation quality. Evaluation results on four datasets show that Sparse RAG can be used to strike an optimal balance between generation quality and computational efficiency, demonstrating its generalizability across tasks.

3325. {τ}-bench: A Benchmark for \underline{T}ool-\underline{A}gent-\underline{U}ser Interaction in Real-World Domains

链接：https://iclr.cc/virtual/2025/poster/28170 abstract： Existing benchmarks for language agents do not set them up to interact with human users or follow domain-specific rules, both of which are vital to safe and realistic deployment. We propose τ-bench, a benchmark with two domains (retail and airline) emulating dynamic conversations between a user (simulated by language models) and a customer service agent provided with domain-specific API tools and policy guidelines. We employ a efficient and faithful evaluation process that compares the database state at the end of a conversation with the annotated goal state, and propose a new metric (pass^k) to evaluate the reliability of agent behavior over multiple trials. Our experiments show that even state-of-the-art function calling agents (gpt-4o) succeed on $<50\%$ of the tasks, and are terribly inconsistent (pass^8 < 25\% in retail). Our findings point to the need for methods that can improve the ability of agents to act consistently and reliably follow rules.

3326. QuaDiM: A Conditional Diffusion Model For Quantum State Property Estimation

链接：https://iclr.cc/virtual/2025/poster/29772 abstract： Quantum state property estimation (QPE) is a fundamental challenge in quantum many-body problems in physics and chemistry, involving the prediction of characteristics such as correlation and entanglement entropy through statistical analysis of quantum measurement data. Recent advances in deep learning have provided powerful solutions, predominantly using auto-regressive models. These models generally assume an intrinsic ordering among qubits, aiming to approximate the classical probability distribution through sequential training. However, unlike natural language, the entanglement structure of qubits lacks an inherent ordering, hurting the motivation of such models. In this paper, we introduce a novel, non-autoregressive generative model called \textbf{\model}, designed for \underline{\textbf{Qua}}ntum state property estimation using \underline{\textbf{Di}}ffusion \underline{\textbf{M}}odels. \model progressively denoises Gaussian noise into the distribution corresponding to the quantum state, encouraging equal, unbiased treatment of all qubits. \model learns to map physical variables to properties of the ground state of the parameterized Hamiltonian during offline training. Afterwards one can sample from the learned distribution conditioned on previously unseen physical variables to collect measurement records and employ post-processing to predict properties of unknown quantum states. We evaluate \model on large-scale QPE tasks using classically simulated data on the 1D anti-ferromagnetic Heisenberg model with the system size up to 100 qubits. Numerical results demonstrate that \model outperforms baseline models, particularly auto-regressive approaches, under conditions of limited measurement data during training and reduced sample complexity during inference.

3327. Learning Interleaved Image-Text Comprehension in Vision-Language Large Models

链接：https://iclr.cc/virtual/2025/poster/28632 abstract： The swift progress of Multi-modal Large Models (MLLMs) has showcased their impressive ability to tackle tasks blending vision and language.Yet, most current models and benchmarks cater to scenarios with a narrow scope of visual and textual contexts.These models often fall short when faced with complex comprehension tasks, which involve navigating through a plethora of irrelevant and potentially misleading information in both text and image forms.To bridge this gap, we introduce a new, more demanding task known as Interleaved Image-Text Comprehension (IITC).This task challenges models to discern and disregard superfluous elements in both images and text to accurately answer questions and to follow intricate instructions to pinpoint the relevant image.In support of this task, we further craft a new VEGA dataset, tailored for the IITC task on scientific content, and devised a subtask, Image-Text Association (ITA), to refine image-text correlation skills.Our evaluation of four leading closed-source models, as well as various open-source models using VEGA, underscores the rigorous nature of IITC.Even the most advanced models, such as Gemini-1.5-pro and GPT4V, only achieved modest success.By employing a multi-task, multi-scale post-training strategy, we have set a robust baseline for MLLMs on the IITC task, attaining an $85.8\%$ accuracy rate in image association and a $0.508$ Rouge score. These results validate the effectiveness of our dataset in improving MLLMs capabilities for nuanced image-text comprehension.

3328. ADIFF: Explaining audio difference using natural language

链接：https://iclr.cc/virtual/2025/poster/28536 abstract：

3329. Taming Transformer Without Using Learning Rate Warmup

链接：https://iclr.cc/virtual/2025/poster/30268 abstract：

3330. Towards a learning theory of representation alignment

链接：https://iclr.cc/virtual/2025/poster/30459 abstract： It has recently been argued that AI models' representations are becoming aligned as their scale and performance increase. Empirical analyses have been designed to support this idea and conjecture the possible alignment of different representations toward a shared statistical model of reality. In this paper, we propose a learning-theoretic perspective to representation alignment. First, we review and connect different notions of alignment based on metric, probabilistic, and spectral ideas. Then, we focus on stitching, a particular approach to understanding the interplay between different representations in the context of a task. Our main contribution here is to relate the properties of stitching to the kernel alignment of the underlying representation. Our results can be seen as a first step toward casting representation alignment as a learning-theoretic problem.

3331. Preference Diffusion for Recommendation

链接：https://iclr.cc/virtual/2025/poster/30900 abstract： Recommender systems aim to predict personalized item rankings by modeling user preference distributions derived from historical behavior data. While diffusion models (DMs) have recently gained attention for their ability to model complex distributions, current DM-based recommenders typically rely on traditional objectives such as mean squared error (MSE) or standard recommendation objectives. These approaches are either suboptimal for personalized ranking tasks or fail to exploit the full generative potential of DMs. To address these limitations, we propose \textbf{PreferDiff}, an optimization objective tailored for DM-based recommenders. PreferDiff reformulates the traditional Bayesian Personalized Ranking (BPR) objective into a log-likelihood generative framework, enabling it to effectively capture user preferences by integrating multiple negative samples. To handle the intractability, we employ variational inference, minimizing the variational upper bound. Furthermore, we replace MSE with cosine error to improve alignment with recommendation tasks, and we balance generative learning and preference modeling to enhance the training stability of DMs. PreferDiff devises three appealing properties. First, it is the first personalized ranking loss designed specifically for DM-based recommenders. Second, it improves ranking performance and accelerates convergence by effectively addressing hard negatives. Third, we establish its theoretical connection to Direct Preference Optimization (DPO), demonstrating its potential to align user preferences within a generative modeling framework. Extensive experiments across six benchmarks validate PreferDiff's superior recommendation performance. Our codes are available at \url{https://github.com/lswhim/PreferDiff}.

3332. Fair Submodular Cover

链接：https://iclr.cc/virtual/2025/poster/29478 abstract： Machine learning algorithms are becoming increasing prevalent in the modern world, and as a result there has been significant recent study into algorithmic fairness in order to minimize the possibility of unintentional bias or discrimination in these algorithms. Submodular optimization problems also arise in many machine learning applications, including those such as data summarization and clustering where fairness is an important concern. In this paper, we initiate the study of the Fair Submodular Cover Problem (FSC). Given a ground set U, a monotone submodular function $f:2^U\to\mathbb{R}_{\ge 0}$, and a threshold τ, the goal of FSC is to find a balanced subset of U with minimum cardinality such that $f(S)\ge\tau$. We first introduce discrete algorithms for FSC that achieve a bicriteria approximation ratio of $(\frac{1}{\varepsilon}, 1-O(\varepsilon))$. We then present a continuous algorithm that achieves a $(\ln\frac{1}{\varepsilon}, 1-O(\varepsilon))$-bicriteria approximation ratio, which matches the best approximation guarantee of submodular cover without a fairness constraint. Finally, we complement our theoretical results with a number of empirical evaluations that demonstrate the efficiency of our algorithms on instances of maximum coverage.

3333. Biologically Plausible Brain Graph Transformer

链接：https://iclr.cc/virtual/2025/poster/28197 abstract： State-of-the-art brain graph analysis methods fail to fully encode the small-world architecture of brain graphs (accompanied by the presence of hubs and functional modules), and therefore lack biological plausibility to some extent. This limitation hinders their ability to accurately represent the brain's structural and functional properties, thereby restricting the effectiveness of machine learning models in tasks such as brain disorder detection. In this work, we propose a novel Biologically Plausible Brain Graph Transformer (BioBGT) that encodes the small-world architecture inherent in brain graphs. Specifically, we present a network entanglement-based node importance encoding technique that captures the structural importance of nodes in global information propagation during brain graph communication, highlighting the biological properties of the brain structure. Furthermore, we introduce a functional module-aware self-attention to preserve the functional segregation and integration characteristics of brain graphs in the learned representations. Experimental results on three benchmark datasets demonstrate that BioBGT outperforms state-of-the-art models, enhancing biologically plausible brain graph representations for various brain graph analytical tasks

3334. Real2Code: Reconstruct Articulated Objects via Code Generation

链接：https://iclr.cc/virtual/2025/poster/30528 abstract： We present Real2Code, a novel approach to reconstructing articulated objects via code generation. Given visual observations of an object, we first reconstruct its part geometry using image segmentation and shape completion. We represent these object parts with oriented bounding boxes, from which a fine-tuned large language model (LLM) predicts joint articulation as code. By leveraging pre-trained vision and language models, our approach scales elegantly with the number of articulated parts, and generalizes from synthetic training data to real world objects in unstructured environments. Experimental results demonstrate that Real2Code significantly outperforms the previous state-of-the-art in terms of reconstruction accuracy, and is the first approach to extrapolate beyond objects' structural complexity in the training set, as we show for objects with up to 10 articulated parts. When incorporated with a stereo reconstruction model, Real2Code moreover generalizes to real-world objects, given only a handful of multi-view RGB images and without the need for depth or camera information.

3335. CertainlyUncertain: A Benchmark and Metric for Multimodal Epistemic and Aleatoric Awareness

链接：https://iclr.cc/virtual/2025/poster/29051 abstract： The ability to acknowledge the inevitable uncertainty in their knowledge and reasoning is a prerequisite for AI systems to be truly truthful and reliable. In this paper, we present a taxonomy of uncertainty specific to vision-language AI systems, distinguishing between epistemic uncertainty (arising from a lack of information) and aleatoric uncertainty (due to inherent unpredictability), and further explore finer categories within. Based on this taxonomy, we synthesize a benchmark dataset, CertainlyUncertain, featuring 178K visual question answering (VQA) samples as contrastive pairs. This is achieved by 1) inpainting images to make previously answerable questions into unanswerable ones; and 2) using image captions to prompt large language models for both answerable and unanswerable questions. Additionally, we introduce a new metric confidence-weighted accuracy, that is well correlated with both accuracy and calibration error, to address the shortcomings of existing metrics. Despite the recent rapid progress in vision-language models (VLMs), evaluations on our benchmark show that they perform poorly in uncertain scenarios. Further experiments demonstrate that supervised fine-tuning with CertainlyUncertain enhances the performance of VLMs, and reduces the calibration error. These improvements extend beyond our benchmark to existing refusal-oriented datasets and show positive results on reducing hallucinations, while maintaining performance on standard VQA benchmarks. Our work underscores the importance of addressing uncertainty in vision-language AI systems to improve their reliability and trustworthiness in real-world applications.

3336. BLEND: Behavior-guided Neural Population Dynamics Modeling via Privileged Knowledge Distillation

链接：https://iclr.cc/virtual/2025/poster/28643 abstract： Modeling the nonlinear dynamics of neuronal populations represents a key pursuit in computational neuroscience. Recent research has increasingly focused on jointly modeling neural activity and behavior to unravel their interconnections. Despite significant efforts, these approaches often necessitate either intricate model designs or oversimplified assumptions. Given the frequent absence of perfectly paired neural-behavioral datasets in real-world scenarios when deploying these models, a critical yet understudied research question emerges: how to develop a model that performs well using only neural activity as input at inference, while benefiting from the insights gained from behavioral signals during training?To this end, we propose BLEND, the Behavior-guided neuraL population dynamics modElling framework via privileged kNowledge Distillation. By considering behavior as privileged information, we train a teacher model that takes both behavior observations (privileged features) and neural activities (regular features) as inputs. A student model is then distilled using only neural activity. Unlike existing methods, our framework is model-agnostic and avoids making strong assumptions about the relationship between behavior and neural activity. This allows BLEND to enhance existing neural dynamics modeling architectures without developing specialized models from scratch. Extensive experiments across neural population activity modeling and transcriptomic neuron identity prediction tasks demonstrate strong capabilities of BLEND, reporting over 50% improvement in behavioral decoding and over 15% improvement in transcriptomic neuron identity prediction after behavior-guided distillation. Furthermore, we empirically explore various behavior-guided distillation strategies within the BLEND framework and present a comprehensive analysis of effectiveness and implications for model performance. Code will be made available at https://github.com/dddavid4real/BLEND.

3337. AI as Humanity's Salieri: Quantifying Linguistic Creativity of Language Models via Systematic Attribution of Machine Text against Web Text

链接：https://iclr.cc/virtual/2025/poster/28676 abstract： Creativity has long been considered one of the most difficult aspect of human intelligence for AI to mimic. However, the rise of Large Language Models (LLMs), like ChatGPT, has raised questions about whether AI can match or even surpass human creativity. We present CREATIVITY INDEX as the first step to quantify the linguistic creativity of a text by reconstructing it from existing text snippets on the web. CREATIVITY INDEX is motivated by the hypothesis that the seemingly remarkable creativity of LLMs may be attributable in large part to the creativity of human-written texts on the web. To compute CREATIVITY INDEX efficiently, we introduce DJ SEARCH, a novel dynamic programming algorithm that can search verbatim and near-verbatim matches of text snippets from a given document against the web. Experiments reveal that the CREATIVITY INDEX of professional human authors is on average 66.2% higher than that of LLMs, and that alignment reduces the CREATIVITY INDEX of LLMs by an average of 30.1%. In addition, we find that distinguished authors like Hemingway exhibit measurably higher CREATIVITY INDEX compared to other human writers. Finally, we demonstrate that CREATIVITY INDEX can be used as a surprisingly effective criterion for zero-shot machine text detection, surpassing the strongest existing zero-shot system, DetectGPT, by a significant margin of 30.2%, and even outperforming the strongest supervised system, GhostBuster, in five out of six domains.

3338. F^3Set: Towards Analyzing Fast, Frequent, and Fine-grained Events from Videos

链接：https://iclr.cc/virtual/2025/poster/32051 abstract： Analyzing Fast, Frequent, and Fine-grained (F^3) events presents a significant challenge in video analytics and multi-modal LLMs. Current methods struggle to identify events that satisfy all the F^3 criteria with high accuracy due to challenges such as motion blur and subtle visual discrepancies. To advance research in video understanding, we introduce F^3Set, a benchmark that consists of video datasets for precise F^3 event detection. Datasets in F^3Set are characterized by their extensive scale and comprehensive detail, usually encompassing over 1,000 event types with precise timestamps and supporting multi-level granularity. Currently, F^3Set contains several sports datasets, and this framework may be extended to other applications as well. We evaluated popular temporal action understanding methods on F^3Set, revealing substantial challenges for existing techniques. Additionally, we propose a new method, F^3ED, for F^3 event detections, achieving superior performance. The dataset, model, and benchmark code are available at https://github.com/F3Set/F3Set.

3339. RecFlow: An Industrial Full Flow Recommendation Dataset

链接：https://iclr.cc/virtual/2025/poster/27909 abstract： Industrial recommendation systems (RS) rely on the multi-stage pipeline to balance effectiveness and efficiency when delivering items from a vast corpus to users. Existing RS benchmark datasets primarily focus on the exposure space, where novel RS algorithms are trained and evaluated. However, when these algorithms transition to real-world industrial RS, they face two critical challenges: (1) handling unexposed items—a significantly larger space than the exposed one, profoundly impacting their practical performance; and (2) overlooking the intricate interplay between multiple stages of the recommendation pipeline, resulting in suboptimal system performance. To bridge the gap between offline RS benchmarks and real-world online environments, we introduce RecFlow—an industrial full-flow recommendation dataset. Unlike existing datasets, RecFlow includes samples not only from the exposure space but also from unexposed items filtered at each stage of the RS funnel. RecFlow comprises 38 million interactions from 42,000 users across nearly 9 million items with additional 1.9 billion stage samples collected from 9.3 million online requests over 37 days and spanning 6 stages. Leveraging RecFlow, we conduct extensive experiments to demonstrate its potential in designing novel algorithms that enhance effectiveness by incorporating stage-specific samples. Some of these algorithms have already been deployed online at KuaiShou, consistently yielding significant gains. We propose RecFlow as the first comprehensive whole-pipeline benchmark dataset for the RS community, enabling research on algorithm design across the entire recommendation pipeline, including selection bias study, debiased algorithms, multi-stage consistency and optimality, multi-task recommendation, and user behavior modeling.

3340. Towards a General Time Series Anomaly Detector with Adaptive Bottlenecks and Dual Adversarial Decoders

链接：https://iclr.cc/virtual/2025/poster/29176 abstract：

3341. LOKI: A Comprehensive Synthetic Data Detection Benchmark using Large Multimodal Models

链接：https://iclr.cc/virtual/2025/poster/27689 abstract：

3342. How Far Are We from True Unlearnability?

链接：https://iclr.cc/virtual/2025/poster/30194 abstract： High-quality data plays an indispensable role in the era of large models, but the use of unauthorized data for model training greatly damages the interests of data owners. To overcome this threat, several unlearnable methods have been proposed, which generate unlearnable examples (UEs) by compromising the training availability of data. Clearly, due to unknown training purposes and the powerful representation learning capabilities of existing models, these data are expected to be unlearnable for various task models, i.e., they will not help improve the model's performance. However, unexpectedly, we find that on the multi-task dataset Taskonomy, UEs still perform well in tasks such as semantic segmentation, failing to exhibit \textit{cross-task unlearnability}. This phenomenon leads us to question: \textit{How far are we from attaining truly unlearnable examples?} We attempt to answer this question from the perspective of model optimization. To this end, we observe the difference in the convergence process between clean and poisoned models using a simple model architecture. Subsequently, from the loss landscape we find that only a part of the critical parameter optimization paths show significant differences, implying a close relationship between the loss landscape and unlearnability. Consequently, we employ the loss landscape to explain the underlying reasons for UEs and propose Sharpness-Aware Learnability (SAL) to quantify the unlearnability of parameters based on this explanation. Furthermore, we propose an Unlearnable Distance (UD) to measure the unlearnability of data based on the SAL distribution of parameters in clean and poisoned models. Finally, we conduct benchmark tests on mainstream unlearnable methods using the proposed UD, aiming to promote community awareness of the capability boundaries of existing unlearnable methods.

3343. Hallo2: Long-Duration and High-Resolution Audio-Driven Portrait Image Animation

链接：https://iclr.cc/virtual/2025/poster/28174 abstract： Recent advances in latent diffusion-based generative models for portrait image animation, such as Hallo, have achieved impressive results in short-duration video synthesis. In this paper, we present updates to Hallo, introducing several design enhancements to extend its capabilities.First, we extend the method to produce long-duration videos. To address substantial challenges such as appearance drift and temporal artifacts, we investigate augmentation strategies within the image space of conditional motion frames. Specifically, we introduce a patch-drop technique augmented with Gaussian noise to enhance visual consistency and temporal coherence over long duration.Second, we achieve 4K resolution portrait video generation. To accomplish this, we implement vector quantization of latent codes and apply temporal alignment techniques to maintain coherence across the temporal dimension. By integrating a high-quality decoder, we realize visual synthesis at 4K resolution.Third, we incorporate adjustable semantic textual labels for portrait expressions as conditional inputs. This extends beyond traditional audio cues to improve controllability and increase the diversity of the generated content. To the best of our knowledge, Hallo2, proposed in this paper, is the first method to achieve 4K resolution and generate hour-long, audio-driven portrait image animations enhanced with textual prompts. We have conducted extensive experiments to evaluate our method on publicly available datasets, including HDTF, CelebV, and our introduced "Wild" dataset. The experimental results demonstrate that our approach achieves state-of-the-art performance in long-duration portrait video animation, successfully generating rich and controllable content at 4K resolution for duration extending up to tens of minutes.

3344. GaussianBlock: Building Part-Aware Compositional and Editable 3D Scene by Primitives and Gaussians

链接：https://iclr.cc/virtual/2025/poster/27818 abstract： Recently, with the development of Neural Radiance Fields and Gaussian Splatting, 3D reconstruction techniques have achieved remarkably high fidelity. However, the latent representations learnt by these methods are highly entangled and lack interpretability. In this paper, we propose a novel part-aware compositional reconstruction method, called GaussianBlock, that enables semantically coherent and disentangled representations, allowing for precise and physical editing akin to building blocks, while simultaneously maintaining high fidelity.Our GaussianBlock introduces a hybrid representation that leverages the advantages of both primitives, known for their flexible actionability and editability, and 3D Gaussians, which excel in reconstruction quality. Specifically, we achieve semantically coherent primitives through a novel attention-guided centering loss derived from 2D semantic priors, complemented by a dynamic splitting and fusion strategy. Furthermore, we utilize 3D Gaussians that hybridize with primitives to refine structural details and enhance fidelity. Additionally, a binding inheritance strategy is employed to strengthen and maintain the connection between the two. Our reconstructed scenes are evidenced to be disentangled, compositional, and compact across diverse benchmarks, enabling seamless, direct and precise editing while maintaining high quality.

3345. Syntactic and Semantic Control of Large Language Models via Sequential Monte Carlo

链接：https://iclr.cc/virtual/2025/poster/27761 abstract： A wide range of LM applications require generating text that conforms to syntactic or semantic constraints. Imposing such constraints can be naturally framed as *probabilistic conditioning*, but exact generation from the resulting distribution—which can differ substantially from the LM's base distribution—is generally intractable. In this work,we develop an architecture for controlled LM generation based on sequential Monte Carlo (SMC). Our SMC framework allows us to flexibly incorporate domain- and problem-specific constraints at inference time, and efficiently reallocate computational resources in light of new information during the course of generation. By comparing to a number of alternatives and ablations on four challenging domains---Python code generation for data science, text-to-SQL, goal inference, and molecule synthesis—we demonstrate that, with little overhead, our approach allows small open-source language models to outperform models over 8$\times$ larger, as well as closed-source, fine-tuned ones. In support of the probabilistic perspective, we show that these performance improvements are driven by better approximation to the posterior distribution. Our system builds on the framework of Lew et al. (2023) and integrates with its *language model probabilistic programming language*, giving users a simple, programmable way to apply SMC to a broad variety of controlled generation problems.

3346. MoDeGPT: Modular Decomposition for Large Language Model Compression

链接：https://iclr.cc/virtual/2025/poster/30778 abstract： Large Language Models (LLMs) have significantly advanced AI with their exceptional performance across a wide range of tasks. However, their extensive computational requirements restrict their use on devices with limited resources.While recent compression methods based on low-rank matrices show potentialsolutions, they often suffer from significant loss of accuracy or introduce substantialoverhead in parameters and inference time. In this paper, we introduce Modular De-composition (MoDeGPT), a new, efficient, and structured compression frameworkthat overcomes these limitations. MoDeGPT jointly decomposes pairs of consecu-tive subcomponents within Transformer blocks, reduces hidden dimensions throughoutput reconstruction on a larger structural scale than conventional low-rank meth-ods, and repurposes three classical matrix decomposition algorithms—Nyströmapproximation, CR decomposition, and SVD—to ensure bounded errors in ournovel decomposition approach. Our experiments show that MoDeGPT, withoutrelying on backward propagation, consistently matches or surpasses the performance of prior techniques that depend on gradient information, while achieving a98% reduction in compute costs when compressing a 13B-parameter model. OnLLaMA-2/3 and OPT models, MoDeGPT retains 90-95% of zero-shot performancewith compression rates of 25-30%. The compression process can be completed ona single GPU in a few hours, boosting inference throughput by up to 46%.

3347. Synthesizing Programmatic Reinforcement Learning Policies with Large Language Model Guided Search

链接：https://iclr.cc/virtual/2025/poster/30780 abstract： Programmatic reinforcement learning (PRL) has been explored for representing policies through programs as a means to achieve interpretability and generalization. Despite promising outcomes, current state-of-the-art PRL methods are hindered by sample inefficiency, necessitating tens of millions of program-environment interactions. To tackle this challenge, we introduce a novel LLM-guided search framework (LLM-GS). Our key insight is to leverage the programming expertise and common sense reasoning of LLMs to enhance the efficiency of assumption-free, random-guessing search methods. We address the challenge of LLMs' inability to generate precise and grammatically correct programs in domain-specific languages (DSLs) by proposing a Pythonic-DSL strategy — an LLM is instructed to initially generate Python codes and then convert them into DSL programs. To further optimize the LLM-generated programs, we develop a search algorithm named Scheduled Hill Climbing, designed to efficiently explore the programmatic search space to improve the programs consistently. Experimental results in the Karel domain demonstrate our LLM-GS framework's superior effectiveness and efficiency. Extensive ablation studies further verify the critical role of our Pythonic-DSL strategy and Scheduled Hill Climbing algorithm. Moreover, we conduct experiments with two novel tasks, showing that LLM-GS enables users without programming skills and knowledge of the domain or DSL to describe the tasks in natural language to obtain performant programs.

3348. Reflexive Guidance: Improving OoDD in Vision-Language Models via Self-Guided Image-Adaptive Concept Generation

链接：https://iclr.cc/virtual/2025/poster/29663 abstract： With the recent emergence of foundation models trained on internet-scale data and demonstrating remarkable generalization capabilities, such foundation models have become more widely adopted, leading to an expanding range of application domains. Despite this rapid proliferation, the trustworthiness of foundation models remains underexplored. Specifically, the out-of-distribution detection (OoDD) capabilities of large vision-language models (LVLMs), such as GPT-4o, which are trained on massive multi-modal data, have not been sufficiently addressed. The disparity between their demonstrated potential and practical reliability raises concerns regarding the safe and trustworthy deployment of foundation models. To address this gap, we evaluate and analyze the OoDD capabilities of various proprietary and open-source LVLMs. Our investigation contributes to a better understanding of how these foundation models represent confidence scores through their generated natural language responses. Furthermore, we propose a self-guided prompting approach, termed Reflexive Guidance (ReGuide), aimed at enhancing the OoDD capability of LVLMs by leveraging self-generated image-adaptive concept suggestions. Experimental results demonstrate that our ReGuide enhances the performance of current LVLMs in both image classification and OoDD tasks. The lists of sampled images, along with the prompts and responses for each sample are available at https://github.com/daintlab/ReGuide.

3349. W-PCA Based Gradient-Free Proxy for Efficient Search of Lightweight Language Models

链接：https://iclr.cc/virtual/2025/poster/29249 abstract： The demand for efficient natural language processing (NLP) systems has led to the development of lightweight language models. Previous work in this area has primarily focused on manual design or training-based neural architecture search (NAS) methods. Recently, zero-shot NAS methods have been proposed for evaluating language models without the need for training. However, prevailing approaches to zero-shot NAS often face challenges such as biased evaluation metrics and computational inefficiencies.In this paper, we introduce weight-weighted PCA (W-PCA), a novel zero-shot NAS method specifically tailored for lightweight language models. Our approach utilizes two evaluation proxies: the parameter count and the number of principal components with cumulative contribution exceeding η in the feed-forward neural (FFN) layer. Additionally, by eliminating the need for gradient computations, we optimize the evaluation time, thus enhancing the efficiency of designing and evaluating lightweight language models.We conduct a comparative analysis on the GLUE and SQuAD datasets to evaluate our approach. The results demonstrate that our method significantly reduces training time compared to one-shot NAS methods and achieves higher scores in the testing phase compared to previous state-of-the-art training-based methods. Furthermore, we perform ranking evaluations on a dataset sampled from the FlexiBERT search space. Our approach exhibits superior ranking correlation and further reduces solving time compared to other zero-shot NAS methods that require gradient computation.

3350. CHASE-SQL: Multi-Path Reasoning and Preference Optimized Candidate Selection in Text-to-SQL

链接：https://iclr.cc/virtual/2025/poster/30489 abstract： We present CHASE-SQL, a novel framework addressing large language model (LLM) performance challenges for Text-to-SQL tasks by leveraging multi-agent modeling and test-time compute for improved candidate generation and selection. CHASE-SQL uses LLMs to generate diverse SQL candidates with: (1) a divide-and-conquer approach to break down complex queries, (2) chain-of-thought reasoning based on query execution plans, and (3) instance-aware synthetic example generation for tailored few-shot demonstrations. A selection agent ranks candidates via pairwise comparisons using a fine-tuned binary selection LLM, offering robust performance. This framework improves SQL query quality and diversity, achieving state-of-the-art execution accuracy of 73.0% on the BIRD Text-to-SQL benchmark test set, topping the leaderboard at the time of submission.

3351. A CLIP-Powered Framework for Robust and Generalizable Data Selection

链接：https://iclr.cc/virtual/2025/poster/30679 abstract： Large-scale datasets have been pivotal to the advancements of deep learning models in recent years, but training on such large datasets inevitably incurs substantial storage and computational overhead. Meanwhile, real-world datasets often contain redundant and noisy data, imposing a negative impact on training efficiency and model performance.Data selection has shown promise in identifying the most representative samples from the entire dataset, which aims to minimize the performance gap with reduced training costs. Existing works typically rely on single-modality information to assign importance scores for individual samples, which may lead to inaccurate assessments, especially when dealing with noisy or corrupted samples.To address this limitation, we propose a novel CLIP-powered data selection framework that leverages multimodal information for more robust and generalizable sample selection. Specifically, our framework consists of three key modules—dataset adaptation, sample scoring, and selection optimization—that together harness extensive pre-trained multimodal knowledge to comprehensively assess sample influence and optimize the selection results through multi-objective optimization. Extensive experiments demonstrate that our approach consistently outperforms existing state-of-the-art baselines on various benchmark datasets. Notably, our method effectively removes noisy or damaged samples from the dataset, enabling it to achieve even higher performance with less data. This indicates that it is not only a way to accelerate training but can also improve overall data quality. The implementation is available at https://github.com/Jackbrocp/clip-powered-data-selection.

3352. Flat Reward in Policy Parameter Space Implies Robust Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/31013 abstract：

3353. Revisiting Convolution Architecture in the Realm of DNA Foundation Models

链接：https://iclr.cc/virtual/2025/poster/30599 abstract： In recent years, A variety of methods based on Transformer and state space model (SSM) architectures have been proposed, advancing foundational DNA language models. However, there is a lack of comparison between these recent approaches and the classical architecture—convolutional networks (CNNs)—on foundation model benchmarks.This raises the question: are CNNs truly being surpassed by these recent approaches based on transformer and SSM architectures? In this paper, we develop a simple but well-designed CNN-based method, termed ConvNova. ConvNova identifies and proposes three effective designs: 1) dilated convolutions, 2) gated convolutions, and 3) a dual-branch framework for gating mechanisms. Through extensive empirical experiments, we demonstrate that ConvNova significantly outperforms recent methods on more than half of the tasks across several foundation model benchmarks. For example, in histone-related tasks, ConvNova exceeds the second-best method by an average of 5.8\%, while generally utilizing fewer parameters and enabling faster computation. In addition, the experiments observed findings that may be related to biological characteristics. This indicates that CNNs are still a strong competitor compared to Transformers and SSMs. We anticipate that this work will spark renewed interest in CNN-based methods for DNA foundation models.

3354. Can One Modality Model Synergize Training of Other Modality Models?

链接：https://iclr.cc/virtual/2025/poster/30961 abstract： Learning with multiple modalities has recently demonstrated significant gains in many domains by maximizing the shared information across modalities. However, the current approaches strongly rely on high-quality paired datasets, which allow co-training from the paired labels from different modalities. In this context, we raise a pivotal question: Can a model with one modality synergize the training of other models with the different modalities, even without the paired multimodal labels? Our answer is 'Yes'. As a figurative description, we argue that a writer, i.e., a language model, can promote the training of a painter, i.e., a visual model, even without the paired ground truth of text and image. We theoretically argue that a superior representation can be achieved by the synergy between two different modalities without paired supervision. As proofs of concept, we broadly confirm the considerable performance gains from the synergy among visual, language, and audio models. From a theoretical viewpoint, we first establish a mathematical foundation of the synergy between two different modality models, where each one is trained with its own modality. From a practical viewpoint, our work aims to broaden the scope of multimodal learning to encompass the synergistic usage of single-modality models, relieving a strong limitation of paired supervision. The code is available at https://github.com/johnjaejunlee95/synergistic-multimodal.

3355. BiGR: Harnessing Binary Latent Codes for Image Generation and Improved Visual Representation Capabilities

链接：https://iclr.cc/virtual/2025/poster/31194 abstract： We introduce BiGR, a novel conditional image generation model using compact binary latent codes for generative training, focusing on enhancing both generation and representation capabilities. BiGR is the first conditional generative model that unifies generation and discrimination within the same framework. BiGR features a binary tokenizer, a masked modeling mechanism, and a binary transcoder for binary code prediction. Additionally, we introduce a novel entropy-ordered sampling method to enable efficient image generation. Extensive experiments validate BiGR's superior performance in generation quality, as measured by FID-50k, and representation capabilities, as evidenced by linear-probe accuracy. Moreover, BiGR showcases zero-shot generalization across various vision tasks, enabling applications such as image inpainting, outpainting, editing, interpolation, and enrichment, without the need for structural modifications. Our findings suggest that BiGR unifies generative and discriminative tasks effectively, paving the way for further advancements in the field. We further enable BiGR to perform text-to-image generation, showcasing its potential for broader applications.

3356. Integral Performance Approximation for Continuous-Time Reinforcement Learning Control

链接：https://iclr.cc/virtual/2025/poster/27693 abstract： We introduce integral performance approximation (IPA), a new continuous-time reinforcement learning (CT-RL) control method. It leverages an affine nonlinear dynamic model, which partially captures the dynamics of the physical environment, alongside state-action trajectory data to enable optimal control with great data efficiency and robust control performance. Utilizing Kleinman algorithm structures allows IPA to provide theoretical guarantees of learning convergence, solution optimality, and closed-loop stability. Furthermore, we demonstrate the effectiveness of IPA on three CT-RL environments including hypersonic vehicle (HSV) control, which has additional challenges caused by unstable and nonminimum phase dynamics. As a result, we demonstrate that the IPA method leads to new, SOTA control design and performance in CT-RL.

3357. Does Safety Training of LLMs Generalize to Semantically Related Natural Prompts?

链接：https://iclr.cc/virtual/2025/poster/30001 abstract： Large Language Models (LLMs) are known to be susceptible to crafted adversarial attacks or jailbreaks that lead to the generation of objectionable content despite being aligned to human preferences using safety fine-tuning methods. While the large dimensionality of input token space makes it inevitable to find adversarial prompts that can jailbreak these models, we aim to evaluate whether safety fine-tuned LLMs are safe against natural prompts which are semantically related to toxic seed prompts that elicit safe responses after alignment. We surprisingly find that popular aligned LLMs such as GPT-4 can be compromised using naive prompts that are NOT even crafted with an objective of jailbreaking the model. Furthermore, we empirically show that given a seed prompt that elicits a toxic response from an unaligned model, one can systematically generate several semantically related natural prompts that can jailbreak aligned LLMs.

Towards this, we propose a method of Response Guided Question Augmentation (ReG-QA) to evaluate the generalization of safety aligned LLMs to natural prompts, that first generates several toxic answers given a seed question using an unaligned LLM (Q to A), and further leverages an LLM to generate questions that are likely to produce these answers (A to Q). We interestingly find that safety fine-tuned LLMs such as GPT-4o are vulnerable to producing natural jailbreak questions from unsafe content (without denial) and can thus be used for the latter (A to Q) step. We obtain attack success rates that are comparable to/ better than leading adversarial attack methods on the JailbreakBench leaderboard, while being significantly more stable against defenses such as Smooth-LLM and Synonym Substitution, which are effective against existing all attacks on the leaderboard.

3358. SymmCD: Symmetry-Preserving Crystal Generation with Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/27763 abstract： Generating novel crystalline materials has potential to lead to advancements in fields such as electronics, energy storage, and catalysis. The defining characteristic of crystals is their symmetry, which plays a central role in determining their physical properties. However, existing crystal generation methods either fail to generate materials that display the symmetries of real-world crystals, or simply replicate the symmetry information from examples in a database. To address this limitation, we propose SymmCD, a novel diffusion-based generative model that explicitly incorporates crystallographic symmetry into the generative process. We decompose crystals into two components and learn their joint distribution through diffusion: 1) the asymmetric unit, the smallest subset of the crystal which can generate the whole crystal through symmetry transformations, and; 2) the symmetry transformations needed to be applied to each atom in the asymmetric unit. We also use a novel and interpretable representation for these transformations, enabling generalization across different crystallographic symmetry groups. We showcase the competitive performance of SymmCD on a subset of the Materials Project, obtaining diverse and valid crystals with realistic symmetries and predicted properties.

3359. A Formal Framework for Understanding Length Generalization in Transformers

链接：https://iclr.cc/virtual/2025/poster/29490 abstract： A major challenge for transformers is generalizing to sequences longer than those observed during training. While previous works have empirically shown that transformers can either succeed or fail at length generalization depending on the task, theoretical understanding of this phenomenon remains limited. In this work, we introduce a rigorous theoretical framework to analyze length generalization in causal transformers with learnable absolute positional encodings. In particular, we characterize those functions that are identifiable in the limit from sufficiently long inputs with absolute positional encodings under an idealized inference scheme using a norm-based regularizer. This enables us to prove the possibility of length generalization for a rich family of problems. We experimentally validate the theory as a predictor of success and failure of length generalization across a range of algorithmic and formal language tasks. Our theory not only explains a broad set of empirical observations but also opens the way to provably predicting length generalization capabilities in transformers.

3360. How Much is a Noisy Image Worth? Data Scaling Laws for Ambient Diffusion.

链接：https://iclr.cc/virtual/2025/poster/28244 abstract： The quality of generative models depends on the quality of the data they are trained on. Creating large-scale, high-quality datasets is often expensive and sometimes impossible, e.g.~in certain scientific applications where there is no access to clean data due to physical or instrumentation constraints. Ambient Diffusion and related frameworks train diffusion models with solely corrupted data (which are usually cheaper to acquire) but ambient models significantly underperform models trained on clean data. We study this phenomenon at scale by training more than 80 models on data with different corruption levels across three datasets ranging from $30,000$ to ≈ 1.3M samples. We show that it is impossible, at these sample sizes, to match the performance of models trained on clean data when only training on noisy data. Yet, a combination of a small set of clean data (e.g.~10% of the total dataset) and a large set of highly noisy data suffices to reach the performance of models trained solely on similar-size datasets of clean data, and in particular to achieve near state-of-the-art performance. We provide theoretical evidence for our findings by developing novel sample complexity bounds for learning from Gaussian Mixtures with heterogeneous variances. Our theoretical model suggests that, for large enough datasets, the effective marginal utility of a noisy sample is exponentially worse that of a clean sample. Providing a small set of clean samples can significantly reduce the sample size requirements for noisy data, as we also observe in our experiments.

3361. Efficient Automated Circuit Discovery in Transformers using Contextual Decomposition

链接：https://iclr.cc/virtual/2025/poster/31044 abstract： Automated mechanistic interpretation research has attracted great interest due to its potential to scale explanations of neural network internals to large models. Existing automated circuit discovery work relies on activation patching or its approximations to identify subgraphs in models for specific tasks (circuits). They often suffer from slow runtime, approximation errors, and specific requirements of metrics, such as non-zero gradients.In this work, we introduce contextual decomposition for transformers (CD-T) to build interpretable circuits in large language models. CD-T can produce circuits at any level of abstraction and is the first to efficiently produce circuits as fine-grained as attention heads at specific sequence positions.CD-T is compatible to all transformer types, and requires no training or manually-crafted examples.CD-T consists of a set of mathematical equations to isolate contribution of model features. Through recursively computing contribution of all nodes in a computational graph of a model using CD-T followed by pruning, we are able to reduce circuit discovery runtime from hours to seconds compared to state-of-the-art baselines.On three standard circuit evaluation datasets (indirect object identification, greater-than comparisons, and docstring completion),we demonstrate that CD-T outperforms ACDC and EAP by better recovering the manual circuits with an average of 97% ROC AUC under low runtimes.In addition, we provide evidence that faithfulness of CD-T circuits is not due to random chance by showing our circuits are 80% more faithful than random circuits of up to 60% of the original model size. Finally, we show CD-T circuits are able to perfectly replicate original models' behavior(faithfulness = 1) using fewer nodes than the baselines for all tasks.Our results underscore the great promise of CD-T for efficient automated mechanistic interpretability, paving the way for new insights into the workings of large language models.

3362. A Theoretically-Principled Sparse, Connected, and Rigid Graph Representation of Molecules

链接：https://iclr.cc/virtual/2025/poster/29832 abstract： Graph neural networks (GNNs) -- learn graph representations by exploiting graph's sparsity, connectivity, and symmetries -- have become indispensable for learning geometric data like molecules. However, the most used graphs (e.g., radial cutoff graphs) in molecular modeling lack theoretical guarantees for achieving connectivity and sparsity simultaneously, which are essential for the performance and scalability of GNNs. Furthermore, existing widely used graph construction methods for molecules lack rigidity, limiting GNNs' ability to exploit graph nodes' spatial arrangement. In this paper, we introduce a new hyperparameter-free graph construction of molecules and beyond with sparsity, connectivity, and rigidity guarantees. Remarkably, our method consistently generates connected and sparse graphs with the edge-to-node ratio being bounded above by 3. Our graphs' rigidity guarantees that edge distances and dihedral angles are sufficient to uniquely determine the general spatial arrangements of atoms. We substantiate the effectiveness and efficiency of our proposed graphs in various molecular modeling benchmarks. Code is available at \url{https://github.com/Utah-Math-Data-Science/UnitSphere}.

3363. Learning Generalizable Skills from Offline Multi-Task Data for Multi-Agent Cooperation

链接：https://iclr.cc/virtual/2025/poster/30225 abstract： Learning cooperative multi-agent policy from offline multi-task data that can generalize to unseen tasks with varying numbers of agents and targets is an attractive problem in many scenarios. Although aggregating general behavior patterns among multiple tasks as skills to improve policy transfer is a promising approach, two primary challenges hinder the further advancement of skill learning in offline multi-task MARL. Firstly, extracting general cooperative behaviors from various action sequences as common skills lacks bringing cooperative temporal knowledge into them. Secondly, existing works only involve common skills and can not adaptively choose independent knowledge as task-specific skills in each task for fine-grained action execution. To tackle these challenges, we propose Hierarchical and Separate Skill Discovery (HiSSD), a novel approach for generalizable offline multi-task MARL through skill learning. HiSSD leverages a hierarchical framework that jointly learns common and task-specific skills. The common skills learn cooperative temporal knowledge and enable in-sample exploitation for offline multi-task MARL. The task-specific skills represent the priors of each task and achieve a task-guided fine-grained action execution. To verify the advancement of our method, we conduct experiments on multi-agent MuJoCo and SMAC benchmarks. After training the policy using HiSSD on offline multi-task data, the empirical results show that HiSSD assigns effective cooperative behaviors and obtains superior performance in unseen tasks.

3364. Efficient Perplexity Bound and Ratio Matching in Discrete Diffusion Language Models

链接：https://iclr.cc/virtual/2025/poster/29916 abstract： While continuous diffusion models excel in modeling continuous distributions, their application to categorical data has been less effective. Recent work has shown that ratio-matching through score-entropy within a continuous-time discrete Markov chain (CTMC) framework serves as a competitive alternative to autoregressive models in language modeling.To enhance this framework, we first introduce three new theorems concerning the KL divergence between the data and learned distribution. Our results serve as the discrete counterpart to those established for continuous diffusion models and allow us to derive an improved upper bound of the perplexity. Second, we empirically show that ratio-matching performed by minimizing the denoising cross-entropy between the clean and corrupted data enables models to outperform those utilizing score-entropy with up to 10\% lower perplexity/generative-perplexity, and 15\% faster training steps. To further support our findings, we introduce and evaluate a novel CTMC transition-rate matrix that allows prediction refinement, and derive the analytic expression for its matrix exponential which facilitates the computation of conditional ratios thus enabling efficient training and generation.

3365. Iterative Substructure Extraction for Molecular Relational Learning with Interactive Graph Information Bottleneck

链接：https://iclr.cc/virtual/2025/poster/31059 abstract： Molecular relational learning (MRL) seeks to understand the interaction behaviors between molecules, a pivotal task in domains such as drug discovery and materials science. Recently, extracting core substructures and modeling their interactions have emerged as mainstream approaches within machine learning-assisted methods. However, these methods still exhibit some limitations, such as insufficient consideration of molecular interactions or capturing substructures that include excessive noise, which hampers precise core substructure extraction.To address these challenges, we present an integrated dynamic framework called Iterative Substructure Extraction (ISE). ISE employs the Expectation-Maximization (EM) algorithm for MRL tasks, where the core substructures of interacting molecules are treated as latent variables and model parameters, respectively. Through iterative refinement, ISE gradually narrows the interactions from the entire molecular structures to just the core substructures.Moreover, to ensure the extracted substructures are concise and compact, we propose the Interactive Graph Information Bottleneck (IGIB) theory, which focuses on capturing the most influential yet minimal interactive substructures. In summary, our approach, guided by the IGIB theory, achieves precise substructure extraction within the ISE framework and is encapsulated in the IGIB-ISE}Extensive experiments validate the superiority of our model over state-of-the-art baselines across various tasks in terms of accuracy, generalizability, and interpretability.

3366. VideoGLUE: Video General Understanding Evaluation of Foundation Models

链接：https://iclr.cc/virtual/2025/poster/31464 abstract： We evaluate the video understanding capabilities of existing foundation models (FMs) using a carefully designed experiment protocol consisting of three hallmark tasks (action recognition,temporal localization, and spatiotemporal localization), eight datasets well received by the community, and four adaptation methods tailoring an FM for downstream tasks. Furthermore,we jointly profile FMs’ efficacy and efficiency when adapting to general video understanding tasks using cost measurements during both training and inference. Our main findings areas follows. First, task-specialized models significantly outperform the seven FMs studied in this work, in sharp contrast to what FMs have achieved in natural language and image understanding. Second, video-native FMs, whose pretraining data mainly contains the video modality, are generally better than image-native FMs in classifying motion-rich videos,localizing actions in time, and understanding a video of more than one action. Third, the video-native FMs can perform well on video tasks under light adaptations to downstream tasks (e.g., freezing the FM backbones), while image-native FMs win in full end-to-end finetuning. The first two observations reveal the need and tremendous opportunities to conduct research on video-focused FMs, and the last confirms that both tasks and adaptation methods matter when it comes to the evaluation of FMs. Our code is released under: https://github.com/tensorflow/models/tree/master/official/projects/videoglue

3367. Utility-Directed Conformal Prediction: A Decision-Aware Framework for Actionable Uncertainty Quantification

链接：https://iclr.cc/virtual/2025/poster/28698 abstract： There is increasing interest in ``decision-focused" machine learning methods which train models to account for how their predictions are used in downstream optimization problems. Doing so can often improve performance on subsequent decision problems. However, current methods for uncertainty quantification do not incorporate any information at all about downstream decisions. We develop a framework based on conformal prediction to produce prediction sets that account for a downstream decision loss function, making them more appropriate to inform high-stakes decision-making. Our approach harnesses the strengths of conformal methods—modularity, model-agnosticism, and statistical coverage guarantees—while incorporating downstream decisions and user-specified utility functions. We prove that our methods retain standard coverage guarantees. Empirical evaluation across a range of datasets and utility metrics demonstrates that our methods achieve significantly lower decision loss compared to standard conformal methods. Additionally, we present a real-world use case in healthcare diagnosis, where our method effectively incorporates the hierarchical structure of dermatological diseases. It successfully generates sets with coherent diagnostic meaning, aiding the triage process during dermatology diagnosis and illustrating how our method can ground high-stakes decision-making on external domain knowledge.

3368. Bridging the Semantic Gap Between Text and Table: A Case Study on NL2SQL

链接：https://iclr.cc/virtual/2025/poster/28237 abstract： The rise of Large Language Models (LLMs) has revolutionized numerous domains, yet these models still exhibit weakness in understanding structured tabular data.Although the growing context window promises to accommodate a larger volume of table contents, it does not inherently improve the model's ability to understand the underlying structure and semantics of tabular data.To bridge the semantic gap between Text and Table, we propose TnT, a table-language model that features multimodal table representations to empower LLMs to effectively and efficiently abstract structure-enriched semantics from tabular data. TnT also introduces a scalable and efficient training pipeline, featuring novel self-supervised tasks, to integrate abstract tabular knowledge into the language modality.Extensive experimental results on NL2SQL demonstrate a much better table understanding of TnT, which achieves up to 14.4 higher execution accuracy compared with traditional text-based table representations.

3369. TIPS: Text-Image Pretraining with Spatial awareness

链接：https://iclr.cc/virtual/2025/poster/30452 abstract： While image-text representation learning has become very popular in recent years, existing models tend to lack spatial awareness and have limited direct applicability for dense understanding tasks.For this reason, self-supervised image-only pretraining is still the go-to method for many dense vision applications (e.g. depth estimation, semantic segmentation), despite the lack of explicit supervisory signals.In this paper, we close this gap between image-text and self-supervised learning, by proposing a novel general-purpose image-text model, which can be effectively used off the shelf for dense and global vision tasks.Our method, which we refer to as Text-Image Pretraining with Spatial awareness (TIPS), leverages two simple and effective insights.First, on textual supervision: we reveal that replacing noisy web image captions by synthetically generated textual descriptions boosts dense understanding performance significantly, due to a much richer signal for learning spatially aware representations.We propose an adapted training method that combines noisy and synthetic captions, resulting in improvements across both dense and global understanding tasks.Second, on the learning technique: we propose to combine contrastive image-text learning with self-supervised masked image modeling, to encourage spatial coherence, unlocking substantial enhancements for downstream applications.Building on these two ideas, we scale our model using the transformer architecture, trained on a curated set of public images.Our experiments are conducted on 8 tasks involving 16 datasets in total, demonstrating strong off-the-shelf performance on both dense and global understanding, for several image-only and image-text tasks.Code and models are released at https://github.com/google-deepmind/tips .

3370. GPromptShield: Elevating Resilience in Graph Prompt Tuning Against Adversarial Attacks

链接：https://iclr.cc/virtual/2025/poster/27744 abstract： The paradigm of ``pre-training and prompt-tuning", with its effectiveness and lightweight characteristics, has rapidly spread from the language field to the graph field. Several pioneering studies have designed specialized prompt functions for diverse downstream graph tasks based on various graph pre-training strategies. These prompts concentrate on the compatibility between the pre-training pretext and downstream graph tasks, aiming to bridge the gap between them. However, designing prompts to blindly adapt to downstream tasks based on this concept neglects crucial security issues. By conducting covert attacks on downstream graph data, we find that even when the downstream task data closely matches that of the pre-training tasks, it is still feasible to generate highly misleading prompts using simple deceptive techniques. In this paper, we shift the primary focus of graph prompts from compatibility to vulnerability issues in adversarial attack scenarios. We design a highly extensible shield defense system for the prompts, which enhances their robustness from two perspectives: \textbf{\textit{Direct Handling}} and \textbf{\textit{Indirect Amplification}}. When downstream graph data exhibits unreliable biases, the former directly combats invalid information by adding hybrid multi-defense prompts to the input graph's feature space, while the latter employs a training strategy that circumvents invalid part and amplifies valid part. We provide a theoretical derivation that proves their feasibility, indicating that unbiased prompts exist under certain conditions on unreliable data. Extensive experiments across various scenarios of adversarial attack (including adaptive and non-adaptive attacks) indicate that the prompts within our shield defense system exhibit enhanced resilience and superiority. Our work explores new perspectives in the field of graph prompts, offering a novel option for downstream robust prompt tuning.

3371. Uncovering Overfitting in Large Language Model Editing

链接：https://iclr.cc/virtual/2025/poster/28074 abstract： Knowledge editing has been proposed as an effective method for updating and correcting the internal knowledge of Large Language Models (LLMs). However, existing editing methods often struggle with complex tasks, such as multi-hop reasoning. In this paper, we identify and investigate the phenomenon of Editing Overfit, where edited models assign disproportionately high probabilities to the edit target, hindering the generalization of new knowledge in complex scenarios. We attribute this issue to the current editing paradigm, which places excessive emphasis on the direct correspondence between the input prompt and the edit target for each edit sample. To further explore this issue, we introduce a new benchmark, EVOKE (EValuation of Editing Overfit in Knowledge Editing), along with fine-grained evaluation metrics. Through comprehensive experiments and analysis, we demonstrate that Editing Overfit is prevalent in current editing methods and that common overfitting mitigation strategies are ineffective in knowledge editing. To overcome this, inspired by LLMs' knowledge recall mechanisms, we propose a new plug-and-play strategy called Learn the Inference (LTI), which introduce a Multi-stage Inference Constraint module to guide the edited models in recalling new knowledge similarly to how unedited LLMs leverage knowledge through in-context learning. Extensive experimental results across a wide range of tasks validate the effectiveness of LTI in mitigating Editing Overfit.

3372. UniDrive: Towards Universal Driving Perception Across Camera Configurations

链接：https://iclr.cc/virtual/2025/poster/28636 abstract： Vision-centric autonomous driving has demonstrated excellent performance with economical sensors. As the fundamental step, 3D perception aims to infer 3D information from 2D images based on 3D-2D projection. This makes driving perception models susceptible to sensor configuration (e.g., camera intrinsics and extrinsics) variations. However, generalizing across camera configurations is important for deploying autonomous driving models on different car models. In this paper, we present UniDrive, a novel framework for vision-centric autonomous driving to achieve universal perception across camera configurations. We deploy a set of unified virtual cameras and propose a ground-aware projection method to effectively transform the original images into these unified virtual views. We further propose a virtual configuration optimization method by minimizing the expected projection error between original cameras and virtual cameras. The proposed virtual camera projection can be applied to existing 3D perception methods as a plug-and-play module to mitigate the challenges posed by camera parameter variability, resulting in more adaptable and reliable driving perception models. To evaluate the effectiveness of our framework, we collect a dataset on CARLA by driving the same routes while only modifying the camera configurations. Experimental results demonstrate that our method trained on one specific camera configuration can generalize to varying configurations with minor performance degradation.

3373. Context-aware Dynamic Pruning for Speech Foundation Models

链接：https://iclr.cc/virtual/2025/poster/28001 abstract： Foundation models, such as large language models, have achieved remarkable success in natural language processing and are evolving into models capable of handling multiple modalities.Listening ability, in particular, is crucial for many applications, leading to research on building speech foundation models. However, the high computational cost of these large models presents a significant challenge for real-world applications. Although substantial efforts have been made to reduce computational costs, such as through pruning techniques, the majority of these approaches are applied primarily during the training phase for specific downstream tasks. In this study, we hypothesize that optimal pruned networks may vary based on contextual factors such as speaker characteristics, languages, and tasks. To address this, we propose a dynamic pruning technique that adapts to these contexts during inference without altering the underlying model. We demonstrated that we could successfully reduce inference time by approximately 30\% while maintaining accuracy in multilingual/multi-task scenarios. We also found that the obtained pruned structure offers meaningful interpretations based on the context, e.g., task-related information emerging as the dominant factor for efficient pruning.

3374. ParetoFlow: Guided Flows in Multi-Objective Optimization

链接：https://iclr.cc/virtual/2025/poster/28472 abstract： In offline multi-objective optimization (MOO), we leverage an offline dataset of designs and their associated labels to simultaneously minimize multiple objectives. This setting more closely mirrors complex real-world problems compared to single-objective optimization. Recent works mainly employ evolutionary algorithms and Bayesian optimization, with limited attention given to the generative modeling capabilities inherent in such data. In this study, we explore generative modeling in offline MOO through flow matching, noted for its effectiveness and efficiency. We introduce \textit{ParetoFlow}, specifically designed to guide flow sampling to approximate the Pareto front. Traditional predictor~(classifier) guidance is inadequate for this purpose because it models only a single objective. In response, we propose a \textit{multi-objective predictor guidance} module that assigns each sample a weight vector, representing a weighted distribution across multiple objective predictions. A local filtering scheme is introduced to address non-convex Pareto fronts. These weights uniformly cover the entire objective space, effectively directing sample generation towards the Pareto front. Since distributions with similar weights tend to generate similar samples, we introduce a \textit{neighboring evolution} module to foster knowledge sharing among neighboring distributions. This module generates offspring from these distributions, and selects the most promising one for the next iteration. Our method achieves state-of-the-art performance across various tasks. Our code is available.

3375. High-Dimensional Bayesian Optimisation with Gaussian Process Prior Variational Autoencoders

链接：https://iclr.cc/virtual/2025/poster/29615 abstract： Bayesian optimisation (BO) using a Gaussian process (GP)-based surrogate model is a powerful tool for solving black-box optimisation problems but does not scale well to high-dimensional data. Previous works have proposed to use variational autoencoders (VAEs) to project high-dimensional data onto a low-dimensional latent space and to implement BO in the inferred latent space. In this work, we propose a conditional generative model for efficient high-dimensional BO that uses a GP surrogate model together with GP prior VAEs. A GP prior VAE extends the standard VAE by conditioning the generative and inference model on auxiliary covariates, capturing complex correlations across samples with a GP. Our model incorporates the observed target quantity values as auxiliary covariates learning a structured latent space that is better suited for the GP-based BO surrogate model. It handles partially observed auxiliary covariates using a unifying probabilistic framework and can also incorporate additional auxiliary covariates that may be available in real-world applications. We demonstrate that our method improves upon existing latent space BO methods on simulated datasets as well as on commonly used benchmarks.

3376. Autonomous Evaluation of LLMs for Truth Maintenance and Reasoning Tasks

链接：https://iclr.cc/virtual/2025/poster/28665 abstract： This paper presents AutoEval, a novel benchmark for scaling Large Language Model (LLM) assessment in formal tasks with clear notions of correctness, such as truth maintenance in translation and logical reasoning. AutoEval is the first benchmarking paradigm that offers several key advantages necessary for scaling objective evaluation of LLMs without human labeling: (a) ability to evaluate LLMs of increasing sophistication by auto-generating tasks at different levels of difficulty; (b) auto-generation of ground truth that eliminates dependence on expensive and time-consuming human annotation; (c) the use of automatically generated, randomized datasets that mitigate the ability of successive LLMs to overfit to static datasets used in many contemporary benchmarks. Empirical analysis shows that an LLM's performance on AutoEval is highly indicative of its performance on a diverse array of other benchmarks focusing on translation and reasoning tasks, making it a valuable autonomous evaluation paradigm in settings where hand-curated datasets can be hard to obtain and/or update.

3377. TimeMixer++: A General Time Series Pattern Machine for Universal Predictive Analysis

链接：https://iclr.cc/virtual/2025/poster/31219 abstract： Time series analysis plays a critical role in numerous applications, supporting tasks such as forecasting, classification, anomaly detection, and imputation. In this work, we present the time series pattern machine (TSPM), a model designed to excel in a broad range of time series tasks through powerful representation and pattern extraction capabilities. Traditional time series models often struggle to capture universal patterns, limiting their effectiveness across diverse tasks. To address this, we define multiple scales in the time domain and various resolutions in the frequency domain, employing various mixing strategies to extract intricate, task-adaptive time series patterns. Specifically, we introduce TimeMixer++, a general-purpose TSPM that processes multi-scale time series using (1) multi-resolution time imaging (MRTI), (2) time image decomposition (TID), (3) multi-scale mixing (MCM), and (4) multi-resolution mixing (MRM) to extract comprehensive temporal patterns. MRTI transforms multi-scale time series into multi-resolution time images, capturing patterns across both temporal and frequency domains. TID leverages dual-axis attention to extract seasonal and trend patterns, while MCM hierarchically aggregates these patterns across scales. MRM adaptively integrates all representations across resolutions. TimeMixer++ achieves state-of-the-art performance across 8 time series analytical tasks, consistently surpassing both general-purpose and task-specific models. Our work marks a promising step toward the next generation of TSPMs, paving the way for further advancements in time series analysis.

3378. Ambient Diffusion Posterior Sampling: Solving Inverse Problems with Diffusion Models Trained on Corrupted Data

链接：https://iclr.cc/virtual/2025/poster/28241 abstract： We provide a framework for solving inverse problems with diffusion models learned from linearly corrupted data. Firstly, we extend the Ambient Diffusion framework to enable training directly from measurements corrupted in the Fourier domain. Subsequently, we train diffusion models for MRI with access only to Fourier subsampled multi-coil measurements at acceleration factors R$=2, 4, 6, 8$. Secondly, we propose $\textit{Ambient Diffusion Posterior Sampling}$ (A-DPS), a reconstruction algorithm that leverages generative models pre-trained on one type of corruption (e.g. image inpainting) to perform posterior sampling on measurements from a different forward process (e.g. image blurring). For MRI reconstruction in high acceleration regimes, we observe that A-DPS models trained on subsampled data are better suited to solving inverse problems than models trained on fully sampled data. We also test the efficacy of A-DPS on natural image datasets (CelebA, FFHQ, and AFHQ) and show that A-DPS can sometimes outperform models trained on clean data for several image restoration tasks in both speed and performance.

3379. HyperDAS: Towards Automating Mechanistic Interpretability with Hypernetworks

链接：https://iclr.cc/virtual/2025/poster/30872 abstract： Mechanistic interpretability has made great strides in identifying neural network features (e.g., directions in hidden activation space) that mediate concepts (e.g., the birth year of a Nobel laureate) and enable predictable manipulation. Distributed alignment search (DAS) leverages supervision from counterfactual data to learn concept features within hidden states, but DAS assumes we can afford to conduct a brute force search over potential feature locations. To address this, we present HyperDAS, a transformer-based hypernetwork architecture that (1) automatically locates the token-positions of the residual stream that a concept is realized in and (2) learns features of those residual stream vectors for the concept. In experiments with Llama3-8B, HyperDAS achieves state-of-the-art performance on the RAVEL benchmark for disentangling concepts in hidden states. In addition, we review the design decisions we made to mitigate the concern that HyperDAS (like all powerful interpretabilty methods) might inject new information into the target model rather than faithfully interpreting it.

3380. Beyond Model Collapse: Scaling Up with Synthesized Data Requires Verification

链接：https://iclr.cc/virtual/2025/poster/29933 abstract： Large Language Models (LLM) are increasingly trained on data generated by other LLMs, either because generated text and images become part of the pre-training corpus, or because synthetized data is used as a replacement for expensive human-annotation. This raises concerns about model collapse, a drop in model performance when their training sets include generated data. Considering that it is easier for both humans and machines to tell between good and bad examples than to generate high-quality samples, we investigate the use of verification on synthesized data to prevent model collapse. We provide a theoretical characterization using Gaussian mixtures, linear classifiers, and linear verifiers to derive conditions with measurable proxies to assess whether the verifier can effectively select synthesized data that leads to optimal performance. We experiment with two practical tasks -- computing matrix eigenvalues with transformers and news summarization with LLMs -- which both exhibit model collapse when trained on generated data, and show that verifiers, even imperfect ones, can indeed be harnessed to prevent model collapse and that our proposed proxy measure strongly correlates with performance.

3381. Online Reinforcement Learning in Non-Stationary Context-Driven Environments

链接：https://iclr.cc/virtual/2025/poster/28535 abstract： We study online reinforcement learning (RL) in non-stationary environments, where a time-varying exogenous context process affects the environment dynamics. Online RL is challenging in such environments due to "catastrophic forgetting" (CF). The agent tends to forget prior knowledge as it trains on new experiences. Prior approaches to mitigate this issue assume task labels (which are often not available in practice), employ brittle regularization heuristics, or use off-policy methods that suffer from instability and poor performance.We present Locally Constrained Policy Optimization (LCPO), an online RL approach that combats CF by anchoring policy outputs on old experiences while optimizing the return on current experiences. To perform this anchoring, LCPO locally constrains policy optimization using samples from experiences that lie outside of the current context distribution. We evaluate LCPO in Mujoco, classic control and computer systems environments with a variety of synthetic and real context traces, and find that it outperforms a variety of baselines in the non-stationary setting, while achieving results on-par with a "prescient" agent trained offline across all context traces.LCPO's source code is available at https://github.com/pouyahmdn/LCPO.

3382. Strong Model Collapse

链接：https://iclr.cc/virtual/2025/poster/32070 abstract： Within the scaling laws paradigm, which underpins the training of large neural networks like ChatGPT and Llama, we consider a supervised regression setting and establish a strong form of the model collapse phenomenon, a critical performance degradation due to synthetic data in the training corpus. Our results show that even the smallest fraction of synthetic data (e.g., as little as 1 per 1000) can still lead to model collapse: larger and larger training sets do not enhance performance. We further investigate whether increasing model size, an approach aligned with current trends in training large language models, exacerbates or mitigates model collapse. In a simplified regime where neural networks are approximated via random projections of tunable size, we both theoretically and empirically show that larger models can amplify model collapse. Interestingly, our theory also indicates that, beyond the interpolation threshold (which can be extremely high for very large datasets), larger models may mitigate the collapse, although they do not entirely prevent it. Our theoretical findings are empirically verified through experiments on language models and neural networks for images.

3383. Optimal Brain Apoptosis

链接：https://iclr.cc/virtual/2025/poster/30781 abstract： The increasing complexity and parameter count of Convolutional Neural Networks (CNNs) and Transformers pose challenges in terms of computational efficiency and resource demands. Pruning has been identified as an effective strategy to address these challenges by removing redundant elements such as neurons, channels, or connections, thereby enhancing computational efficiency without heavily compromising performance. This paper builds on the foundational work of Optimal Brain Damage (OBD) by advancing the methodology of parameter importance estimation using the Hessian matrix. Unlike previous approaches that rely on approximations, we introduce Optimal Brain Apoptosis (OBA), a novel pruning method that calculates the Hessian-vector product value directly for each parameter. By decomposing the Hessian matrix across network layers and identifying conditions under which inter-layer Hessian submatrices are non-zero, we propose a highly efficient technique for computing the second-order Taylor expansion of parameters. This approach allows for a more precise pruning process, particularly in the context of CNNs and Transformers, as validated in our experiments including VGG19, ResNet32, ResNet50, and ViT-B/16 on CIFAR10, CIFAR100 and Imagenet datasets. Our code is available at https://github.com/NEU-REAL/OBA.

3384. Node-Time Conditional Prompt Learning in Dynamic Graphs

链接：https://iclr.cc/virtual/2025/poster/28577 abstract： Dynamic graphs capture evolving interactions between entities, such as in social networks, online learning platforms, and crowdsourcing projects. For dynamic graph modeling, dynamic graph neural networks (DGNNs) have emerged as a mainstream technique. However, they are generally pre-trained on the link prediction task, leaving a significant gap from the objectives of downstream tasks such as node classification. To bridge the gap, prompt-based learning has gained traction on graphs, but most existing efforts focus on static graphs, neglecting the evolution of dynamic graphs. In this paper, we propose DyGPrompt, a novel pre-training and prompt learning framework for dynamic graph modeling. First, we design dual prompts to address the gap in both task objectives and temporal variations across pre-training and downstream tasks. Second, we recognize that node and time patterns often characterize each other, and propose dual condition-nets to model the evolving node-time patterns in downstream tasks. Finally, we thoroughly evaluate and analyze DyGPrompt through extensive experiments on four public datasets.

3385. Autoregressive Video Generation without Vector Quantization

链接：https://iclr.cc/virtual/2025/poster/30117 abstract： This paper presents a novel approach that enables autoregressive video generation with high efficiency. We propose to reformulate the video generation problem as a non-quantized autoregressive modeling of temporal frame-by-frame prediction and spatial set-by-set prediction. Unlike raster-scan prediction in prior autoregressive models or joint distribution modeling of fixed-length tokens in diffusion models, our approach maintains the causal property of GPT-style models for flexible in-context capabilities, while leveraging bidirectional modeling within individual frames for efficiency. With the proposed approach, we train a novel video autoregressive model without vector quantization, termed NOVA. Our results demonstrate that NOVA surpasses prior autoregressive video models in data efficiency, inference speed, visual fidelity, and video fluency, even with a much smaller model capacity, i.e., 0.6B parameters. NOVA also outperforms state-of-the-art image diffusion models in text-to-image generation tasks, with a significantly lower training cost. Additionally, NOVA generalizes well across extended video durations and enables diverse zero-shot applications in one unified model. Code and models are publicly available at https://github.com/baaivision/NOVA.

3386. A Stochastic Approach to the Subset Selection Problem via Mirror

Descent

链接：https://iclr.cc/virtual/2025/poster/30952 abstract： The subset selection problem is fundamental in machine learning and other fields of computer science.We introduce a stochastic formulation for the minimum cost subset selection problem in a black box setting, in which only the subset metric value is available.Subsequently, we can handle two-stage schemes, with an outer subset-selection component and an inner subset cost evaluation component. We propose formulating the subset selection problem in a stochastic manner by choosing subsets at random from a distribution whose parameters are learned. Two stochastic formulations are proposed.The first explicitly restricts the subset's cardinality, and the second yields the desired cardinality in expectation.The distribution is parameterized by a decision variable, which we optimize using Stochastic Mirror Descent.Our choice of distributions yields constructive closed-form unbiased stochastic gradient formulas and convergence guarantees, including a rate with favorable dependency on the problem parameters.Empirical evaluation of selecting a subset of layers in transfer learning complements our theoretical findings and demonstrates the potential benefits of our approach.

3387. GReaTer: Gradients Over Reasoning Makes Smaller Language Models Strong Prompt Optimizers

链接：https://iclr.cc/virtual/2025/poster/28876 abstract： The effectiveness of large language models (LLMs) is closely tied to the design of prompts, making prompt optimization essential for enhancing their performance across a wide range of tasks. Although recent advancements have focused on automating prompt engineering, many existing approaches rely exclusively on textual feedback, refining prompts based solely on inference errors identified by large, computationally expensive LLMs. Unfortunately, smaller models struggle to generate high-quality feedback, resulting in complete dependence on large LLM judgment. Moreover, these methods fail to leverage more direct and finer-grained information, such as gradients, due to operating purely in text space. To this end, we introduce, we introduce GReaTer, a novel prompt optimization technique that directly incorporates gradient information over task-specific reasoning. By utilizing task loss gradients, GReaTer enables self-optimization of prompts for smaller, lightweight language models (LM) without the need for costly closed-source LLMs, while maintaining reasonable prompt structures. This allows high-performance prompt optimization without dependence on massive LLMs, closing the gap between smaller models and the sophisticated reasoning often needed for prompt refinement. Extensive evaluations across diverse tasks demonstrate that \ours consistently outperforms previous methods, even those reliant on powerful LLMs. Additionally, GReaTer-optimized prompts frequently exhibit better transferability and, in some cases, boost task performance to levels comparable to or surpassing those achieved by larger language models, highlighting the effectiveness of "gradient over reasoning"-based prompt optimization. Code of GReaTer is available at: https://github.com/psunlpgroup/GreaTer

3388. Masked Temporal Interpolation Diffusion for Procedure Planning in Instructional Videos

链接：https://iclr.cc/virtual/2025/poster/30210 abstract： In this paper, we address the challenge of procedure planning in instructional videos, aiming to generate coherent and task-aligned action sequences from start and end visual observations. Previous work has mainly relied on text-level supervision to bridge the gap between observed states and unobserved actions, but it struggles with capturing intricate temporal relationships among actions. Building on these efforts, we propose the Masked Temporal Interpolation Diffusion (MTID) model that introduces a latent space temporal interpolation module within the diffusion model. This module leverages a learnable interpolation matrix to generate intermediate latent features, thereby augmenting visual supervision with richer mid-state details. By integrating this enriched supervision into the model, we enable end-to-end training tailored to task-specific requirements, significantly enhancing the model's capacity to predict temporally coherent action sequences. Additionally, we introduce an action-aware mask projection mechanism to restrict the action generation space, combined with a task-adaptive masked proximity loss to prioritize more accurate reasoning results close to the given start and end states over those in intermediate steps. Simultaneously, it filters out task-irrelevant action predictions, leading to contextually aware action sequences. Experimental results across three widely used benchmark datasets demonstrate that our MTID achieves promising action planning performance on most metrics.

3389. DEPT: Decoupled Embeddings for Pre-training Language Models

链接：https://iclr.cc/virtual/2025/poster/27901 abstract： Language Model pre-training uses broad data mixtures to enhance performance across domains and languages. However, training on such heterogeneous text corpora requires extensive and expensive efforts. Since these data sources vary significantly in lexical, syntactic, and semantic aspects, they cause negative interference or the ``curse of multilinguality''. To address these challenges we propose a communication-efficient pre-training framework, DEPT. Our method decouples embeddings from the transformer body while simultaneously training the latter on multiple data sources without requiring a shared vocabulary. DEPT can: (1) train robustly and effectively under significant data heterogeneity, (2) minimize token embedding parameters to only what the data source vocabulary requires, while cutting communication costs in direct proportion to both the communication frequency and the reduction in parameters, (3) enhance transformer body plasticity and generalization, improving both average perplexity (up to 20%) and downstream task performance, and (4) enable training with custom optimized vocabularies per data source. We demonstrate DEPT's potential via the first vocabulary-agnostic federated pre-training of billion-scale models, reducing communication costs by orders of magnitude and embedding memory by 4-5x.

3390. Agent-to-Sim: Learning Interactive Behavior Models from Casual Longitudinal Videos

链接：https://iclr.cc/virtual/2025/poster/27750 abstract：

3391. MAPS: Advancing Multi-Modal Reasoning in Expert-Level Physical Science

链接：https://iclr.cc/virtual/2025/poster/30279 abstract： Pre-trained on extensive text and image corpora, current Multi-Modal Large Language Models (MLLM) have shown strong capabilities in general visual reasoning tasks. However, their performance is still lacking in physical domains that require understanding diagrams with complex physical structures and quantitative analysis based on multi-modal information. To address this, we develop a new framework, named Multi-Modal Scientific ReAsoning with Physics Perception and Simulation (MAPS) based on an MLLM. MAPS decomposes expert-level multi-modal reasoning task into physical diagram understanding via a Physical Perception Model (PPM) and reasoning with physical knowledge via a simulator. The PPM module is obtained by fine-tuning a visual language model using carefully designed synthetic data with paired physical diagrams and corresponding simulation language descriptions. At the inference stage, MAPS integrates the simulation language description of the input diagram provided by PPM and results obtained through a Chain-of-Simulation process with MLLM to derive the underlying rationale and the final answer. Validated using our collected college-level circuit analysis problems, MAPS significantly improves reasoning accuracy of MLLM and outperforms all existing models. The results confirm MAPS offers a promising direction for enhancing multi-modal scientific reasoning ability of MLLMs. We will release our code, model and dataset used for our experiments upon publishing of this paper.

3392. Automated Proof Generation for Rust Code via Self-Evolution

链接：https://iclr.cc/virtual/2025/poster/31144 abstract： Ensuring correctness is crucial for code generation. Formal verification offers adefinitive assurance of correctness, but demands substantial human effort in proofconstruction and hence raises a pressing need for automation. The primary obsta-cle lies in the severe lack of data—there is much fewer proofs than code snippetsfor Large Language Models (LLMs) to train upon. In this paper, we introduceSAFE, a framework that overcomes the lack of human-written proofs to enableautomated proof generation of Rust code. SAFE establishes a self-evolving cyclewhere data synthesis and fine-tuning collaborate to enhance the model capability,leveraging the definitive power of a symbolic verifier in telling correct proofs fromincorrect ones. SAFE also re-purposes the large number of synthesized incorrectproofs to train the self-debugging capability of the fine-tuned models, empoweringthem to fix incorrect proofs based on the verifier’s feedback. SAFE demonstratessuperior efficiency and precision compared to GPT-4o. Through tens of thousandsof synthesized proofs and the self-debugging mechanism, we improve the capa-bility of open-source models, initially unacquainted with formal verification, toautomatically write proofs for Rust code. This advancement leads to a signifi-cant improvement in performance, achieving a 52.52% accuracy rate in a bench-mark crafted by human experts, a significant leap over GPT-4o’s performance of14.39%.

3393. Century: A Framework and Dataset for Evaluating Historical Contextualisation of Sensitive Images

链接：https://iclr.cc/virtual/2025/poster/32113 abstract： How do multi-modal generative models describe images of recent historical events and figures, whose legacies may be nuanced, multifaceted, or contested? This task necessitates not only accurate visual recognition, but also socio-cultural knowledge and cross-modal reasoning. To address this evaluation challenge, we introduce Century -- a novel dataset of sensitive historical images. This dataset consists of 1,500 images from recent history, created through an automated method combining knowledge graphs and language models with quality and diversity criteria created from the practices of museums and digital archives. We demonstrate through automated and human evaluation that this method produces a set of images that depict events and figures that are diverse across topics and represents all regions of the world.We additionally propose an evaluation framework for evaluating the historical contextualisation capabilities along dimensions of accuracy, thoroughness, and objectivity. We demonstrate this approach by using Century to evaluate four foundation models, scoring performance using both automated and human evaluation. We find that historical contextualisation of sensitive images poses a significant challenge for modern multi-modal foundation models, and offer practical recommendations for how developers can use Century to evaluate improvements to models and applications.

3394. Bridging Compressed Image Latents and Multimodal Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30277 abstract： This paper presents the first-ever study of adapting compressed image latents to suit the needs of downstream vision tasks that adopt Multimodal Large Language Models (MLLMs). MLLMs have extended the success of large language models to modalities (e.g. images) beyond text, but their billion scale hinders deployment on resource-constrained end devices. While cloud-hosted MLLMs could be available, transmitting raw, uncompressed images captured by end devices to the cloud requires an efficient image compression system. To address this, we focus on emerging neural image compression and propose a novel framework with a lightweight transform-neck and a surrogate loss to adapt compressed image latents for MLLM-based vision tasks. Given the huge scale of MLLMs, our framework excludes the entire downstream MLLM except part of its visual encoder from training our system. This stands out from most existing coding for machine approaches that involve downstream networks in training and thus could be impractical when the networks are MLLMs. The proposed framework is general in that it is applicable to various MLLMs, neural image codecs, and multiple application scenarios, where the neural image codec can be (1) pre-trained for human perception without updating, (2) fully updated for joint human and machine perception, or (3) fully updated for only machine perception. Extensive experiments on different neural image codecs and various MLLMs show that our method achieves great rate-accuracy performance with much less complexity.

3395. Fast Feedforward 3D Gaussian Splatting Compression

链接：https://iclr.cc/virtual/2025/poster/30473 abstract： With 3D Gaussian Splatting (3DGS) advancing real-time and high-fidelity rendering for novel view synthesis, storage requirements pose challenges for their widespread adoption. Although various compression techniques have been proposed, previous art suffers from a common limitation: for any existing 3DGS, per-scene optimization is needed to achieve compression, making the compression sluggish and slow. To address this issue, we introduce Fast Compression of 3D Gaussian Splatting (FCGS), an optimization-free model that can compress 3DGS representations rapidly in a single feed-forward pass, which significantly reduces compression time from minutes to seconds. To enhance compression efficiency, we propose a multi-path entropy module that assigns Gaussian attributes to different entropy constraint paths for balance between size and fidelity. We also carefully design both inter- and intra-Gaussian context models to remove redundancies among the unstructured Gaussian blobs. Overall, FCGS achieves a compression ratio of over 20X while maintaining fidelity, surpassing most per-scene SOTA optimization-based methods. Code: github.com/YihangChen-ee/FCGS.

3396. Selective Label Enhancement Learning for Test-Time Adaptation

链接：https://iclr.cc/virtual/2025/poster/31070 abstract： Test-time adaptation (TTA) aims to adapt a pre-trained model to the target domain using only unlabeled test samples. Most existing TTA approaches rely on definite pseudo-labels, inevitably introducing false labels and failing to capture uncertainty for each test sample. This prevents pseudo-labels from being flexibly refined as the model adapts during training, limiting their potential for performance improvement. To address this, we propose the Progressive Adaptation with Selective Label Enhancement (PASLE) framework. Instead of definite labels, PASLE assigns candidate pseudo-label sets to uncertain ones via selective label enhancement. Specifically, PASLE partitions data into confident/uncertain subsets, assigning one-hot labels to confident samples and candidate sets to uncertain ones. The model progressively trains on certain/uncertain pseudo-labeled data while dynamically refining uncertain pseudo-labels, leveraging increasing target adaptation monitored throughout training. Experiments on various benchmark datasets validate the effectiveness of the proposed approach.

3397. SANER: Annotation-free Societal Attribute Neutralizer for Debiasing CLIP

链接：https://iclr.cc/virtual/2025/poster/27802 abstract： Large-scale vision-language models, such as CLIP, are known to contain societal bias regarding protected attributes (e.g., gender, age). This paper aims to address the problems of societal bias in CLIP. Although previous studies have proposed to debias societal bias through adversarial learning or test-time projecting, our comprehensive study of these works identifies two critical limitations: 1) loss of attribute information when it is explicitly disclosed in the input and 2) use of the attribute annotations during debiasing process. To mitigate societal bias in CLIP and overcome these limitations simultaneously, we introduce a simple-yet-effective debiasing method called SANER (societal attribute neutralizer) that eliminates attribute information from CLIP text features only of attribute-neutral descriptions. Experimental results show that SANER, which does not require attribute annotations and preserves original information for attribute-specific descriptions, demonstrates superior debiasing ability than the existing methods.

3398. Towards Calibrated Deep Clustering Network

链接：https://iclr.cc/virtual/2025/poster/30085 abstract： Deep clustering has exhibited remarkable performance; however, the over confidence problem, i.e., the estimated confidence for a sample belonging to a particular cluster greatly exceeds its actual prediction accuracy, has been over looked in prior research. To tackle this critical issue, we pioneer the development of a calibrated deep clustering framework. Specifically, we propose a novel dualhead (calibration head and clustering head) deep clustering model that can effectively calibrate the estimated confidence and the actual accuracy. The calibration head adjusts the overconfident predictions of the clustering head, generating prediction confidence that matches the model learning status. Then, the clustering head dynamically selects reliable high-confidence samples estimated by the calibration head for pseudo-label self-training. Additionally, we introduce an effective network initialization strategy that enhances both training speed and network robustness. The effectiveness of the proposed calibration approach and initialization strategy are both endorsed with solid theoretical guarantees. Extensive experiments demonstrate the proposed calibrated deep clustering model not only surpasses the state-of-the-art deep clustering methods by 5× on average in terms of expected calibration error, but also significantly outperforms them in terms of clustering accuracy. The code is available at https://github.com/ChengJianH/CDC.

3399. ConMix: Contrastive Mixup at Representation Level for Long-tailed

Deep Clustering

链接：https://iclr.cc/virtual/2025/poster/31057 abstract： Deep clustering has made remarkable progress in recent years. However, most existing deep clustering methods assume that distributions of different clusters are balanced or roughly balanced, which are not consistent with the common long-tailed distributions in reality. In nature, the datasets often follow long-tailed distributions, leading to biased models being trained with significant performance drop. Despite the widespread proposal of many long-tailed learning approaches with supervision information, research on long-tailed deep clustering remains almost uncharted. Unaware of the data distribution and sample labels, long-tailed deep clustering is highly challenging. To tackle this problem, we propose a novel contrastive mixup method for long-tailed deep clustering, named ConMix. The proposed method makes innovations to mixup representations in contrastive learning to enhance deep clustering in long-tailed scenarios. Neural networks trained with ConMix can learn more discriminative representations, thus achieve better long-tailed deep clustering performance. We theoretically prove that ConMix works through re-balancing loss for classes with different long-tailed degree. We evaluate our method on widely used benchmark datasets with different imbalance ratios, suggesting it outperforms many state-of-the-art deep clustering approaches. The code is available at https://github.com/LZX-001/ConMix.

3400. ProAdvPrompter: A Two-Stage Journey to Effective Adversarial Prompting for LLMs

链接：https://iclr.cc/virtual/2025/poster/28021 abstract： As large language models (LLMs) are increasingly being integrated into various real-world applications, the identification of their vulnerabilities to jailbreaking attacks becomes an essential component of ensuring the safety and reliability of LLMs. Previous studies have developed LLM assistants, known as the adversarial prompter, to automatically generate suffixes that manipulate target LLMs into generating harmful and undesirable outputs.However, these approaches often suffer from low performance or generate semantically meaningless prompts, which can be easily identified by perplexity-based defenses.In this paper, we introduce a novel two-stage method, $\texttt{ProAdvPrompter}$, that significantly improves the performance of adversarial prompters.In $\texttt{ProAdvPrompter}$, the first stage (Exploration) utilizes the loss information to guide the adversarial prompter in generating suffixes that are more likely to elicit harmful responses.Then the second stage (Exploitation) iteratively fine-tunes the prompter using high-quality generated adversarial suffixes to further boost performance.Additionally, we incorporate the prompt template to aid in the Exploration stage and propose a filtering mechanism to accelerate the training process in the Exploitation stage.We evaluate $\texttt{ProAdvPrompter}$ against the well-aligned LLMs (i.e., Llama2-Chat-7B and Llama3-chat-8B), achieving attack success rates of 99.68% and 97.12% respectively after 10 trials on the AdvBench dataset, thereby enhancing performance by $\sim 2$ times compared to previous works.Moreover, $\texttt{ProAdvPrompter}$ reduces training time by 20% on Llama3-Instruct-8B, generates more generalized adversarial suffixes, and demonstrates resilience against the perplexity defense.An ablation study further evaluates the effects of key components in $\texttt{ProAdvPrompter}$ (the prompt template and the filtering mechanism).