3401. Growth Inhibitors for Suppressing Inappropriate Image Concepts in Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/27876 abstract： Despite their remarkable image generation capabilities, text-to-image diffusion models inadvertently learn inappropriate concepts from vast and unfiltered training data, which leads to various ethical and business risks. Specifically, model-generated images may exhibit not safe for work (NSFW) content and style copyright infringements. The prompts that result in these problems often do not include explicit unsafe words; instead, they contain obscure and associative terms, which are referred to as implicit unsafe prompts. Existing approaches directly fine-tune models under textual guidance to alter the cognition of the diffusion model, thereby erasing inappropriate concepts. This not only requires concept-specific fine-tuning but may also incur catastrophic forgetting. To address these issues, we explore the representation of inappropriate concepts in the image space and guide them towards more suitable ones by injecting growth inhibitors, which are tailored based on the identified features related to inappropriate concepts during the diffusion process. Additionally, due to the varying degrees and scopes of inappropriate concepts, we train an adapter to infer the corresponding suppression scale during the injection process. Our method effectively captures the manifestation of subtle words at the image level, enabling direct and efficient erasure of target concepts without the need for fine-tuning. Through extensive experimentation, we demonstrate that our approach achieves superior erasure results with little effect on other normal concepts while preserving image quality and semantics.

3402. The OMG dataset: An Open MetaGenomic corpus for mixed-modality genomic language modeling

链接：https://iclr.cc/virtual/2025/poster/28618 abstract： Biological language model performance depends heavily on pretraining data quality, diversity, and size. While metagenomic datasets feature enormous biological diversity, their utilization as pretraining data has been limited due to challenges in data accessibility, quality filtering and deduplication. Here, we present the Open MetaGenomic (OMG) corpus, a genomic pretraining dataset totalling 3.1T base pairs and 3.3B protein coding sequences, obtained by combining two largest metagenomic dataset repositories (JGI's IMG and EMBL's MGnify). We first document the composition of the dataset and describe the quality filtering steps taken to remove poor quality data. We make the OMG corpus available as a mixed-modality genomic sequence dataset that represents multi-gene encoding genomic sequences with translated amino acids for protein coding sequences, and nucleic acids for intergenic sequences. We train the first mixed-modality genomic language model (gLM2) that leverages genomic context information to learn robust functional representations, as well as coevolutionary signals in protein-protein interfaces and genomic regulatory syntax. Furthermore, we show that deduplication in embedding space can be used to balance the corpus, demonstrating improved performance on downstream tasks. The OMG dataset is publicly hosted on the Hugging Face Hub at https://huggingface.co/datasets/tattabio/OMG and gLM2 is available at https://huggingface.co/tattabio/gLM2_650M.

3403. CATCH: Channel-Aware Multivariate Time Series Anomaly Detection via Frequency Patching

链接：https://iclr.cc/virtual/2025/poster/28489 abstract： Anomaly detection in multivariate time series is challenging as heterogeneous subsequence anomalies may occur. Reconstruction-based methods, which focus on learning normal patterns in the frequency domain to detect diverse abnormal subsequences, achieve promising results, while still falling short on capturing fine-grained frequency characteristics and channel correlations. To contend with the limitations, we introduce CATCH, a framework based on frequency patching. We propose to patchify the frequency domain into frequency bands, which enhances its ability to capture fine-grained frequency characteristics. To perceive appropriate channel correlations, we propose a Channel Fusion Module (CFM), which features a patch-wise mask generator and a masked-attention mechanism. Driven by a bi-level multi-objective optimization algorithm, the CFM is encouraged to iteratively discover appropriate patch-wise channel correlations, and to cluster relevant channels while isolating adverse effects from irrelevant channels. Extensive experiments on 10 real-world datasets and 12 synthetic datasets demonstrate that CATCH achieves state-of-the-art performance. We make our code and datasets available at https://github.com/decisionintelligence/CATCH.

3404. Exposure Bracketing Is All You Need For A High-Quality Image

链接：https://iclr.cc/virtual/2025/poster/28212 abstract： It is highly desired but challenging to acquire high-quality photos with clear content in low-light environments. Although multi-image processing methods (using burst, dual-exposure, or multi-exposure images) have made significant progress in addressing this issue, they typically focus on specific restoration or enhancement problems, and do not fully explore the potential of utilizing multiple images. Motivated by the fact that multi-exposure images are complementary in denoising, deblurring, high dynamic range imaging, and super-resolution, we propose to utilize exposure bracketing photography to get a high-quality image by combining these tasks in this work. Due to the difficulty in collecting real-world pairs, we suggest a solution that first pre-trains the model with synthetic paired data and then adapts it to real-world unlabeled images. In particular, a temporally modulated recurrent network (TMRNet) and self-supervised adaptation method are proposed. Moreover, we construct a data simulation pipeline to synthesize pairs and collect real-world images from 200 nighttime scenarios. Experiments on both datasets show that our method performs favorably against the state-of-the-art multi-image processing ones. Code and datasets are available at https://github.com/cszhilu1998/BracketIRE.

3405. GlycanML: A Multi-Task and Multi-Structure Benchmark for Glycan Machine Learning

链接：https://iclr.cc/virtual/2025/poster/28329 abstract： Glycans are basic biomolecules and perform essential functions within living organisms. The rapid increase of functional glycan data provides a good opportunity for machine learning solutions to glycan understanding. However, there still lacks a standard machine learning benchmark for glycan property and function prediction. In this work, we fill this blank by building a comprehensive benchmark for Glycan Machine Learning (GlycanML). The GlycanML benchmark consists of diverse types of tasks including glycan taxonomy prediction, glycan immunogenicity prediction, glycosylation type prediction, and protein-glycan interaction prediction. Glycans can be represented by both sequences and graphs in GlycanML, which enables us to extensively evaluate sequence-based models and graph neural networks (GNNs) on benchmark tasks. Furthermore, by concurrently performing eight glycan taxonomy prediction tasks, we introduce the GlycanML-MTL testbed for multi-task learning (MTL) algorithms. Also, we evaluate how taxonomy prediction can boost other three function prediction tasks by MTL. Experimental results show the superiority of modeling glycans with multi-relational GNNs, and suitable MTL methods can further boost model performance. We provide all datasets and source codes at https://github.com/GlycanML/GlycanML and maintain a leaderboard at https://GlycanML.github.io/project

3406. BlueSuffix: Reinforced Blue Teaming for Vision-Language Models Against Jailbreak Attacks

链接：https://iclr.cc/virtual/2025/poster/27811 abstract： In this paper, we focus on black-box defense for VLMs against jailbreak attacks.Existing black-box defense methods are either unimodal or bimodal. Unimodal methods enhance either the vision or language module of the VLM, while bimodal methods robustify the model through text-image representation realignment. However, these methods suffer from two limitations: 1) they fail to fully exploit the cross-modal information, or 2) they degrade the model performance on benign inputs.To address these limitations, we propose a novel blue-team method BlueSuffix that defends target VLMs against jailbreak attacks without compromising its performance under black-box setting. BlueSuffix includes three key components: 1) a visual purifier against jailbreak images, 2) a textual purifier against jailbreak texts, and 3) a blue-team suffix generator using reinforcement fine-tuning for enhancing cross-modal robustness. We empirically show on four VLMs (LLaVA, MiniGPT-4, InstructionBLIP, and Gemini) and four safety benchmarks (Harmful Instruction, AdvBench, MM-SafetyBench, and RedTeam-2K) that BlueSuffix outperforms the baseline defenses by a significant margin. Our BlueSuffix opens up a promising direction for defending VLMs against jailbreak attacks. Code is available at https://github.com/Vinsonzyh/BlueSuffix.

3407. Scaling Long Context Training Data by Long-Distance Referrals

链接：https://iclr.cc/virtual/2025/poster/28036 abstract： Training large language models for long context understanding faces the challenge of data shortage.Previous data engineering approaches mechanically concatenate short documents, which may create many pseudo long documents but raise concerns about data quality.In this paper, we study the core attribute of high quality data for long context training, and provide a data pipeline, LongPack, to scalesuch data.We found that long distance referrals, which occur in natural long documents, are crucial for long-context training.However, simply concatenating short documents does not reliably generate these relations.We further show that the density of long-distance referrals, which is higher in longer documents, has a key role in training efficiency, making previous upsampling methods suboptimal.To enrich long documents, we propose LongPack, a data pipeline that constructs long documents by packing shorter ones based on referral relationships.Specifically, for web pages, which are the primary source for language model training, we found hyper-link a native signal for such a relation.By packing web pages through their hyper-link connection, we can create longer, high-quality documents.Our experiments demonstrate that LongPackis highly scalable, generating a corpus of long documents equivalent in size to an entire pretraining dataset using just 0.5% root documents.Furthermore, the constructed documents have a ‘near-natural’ quality as innate long documents for long context training, reaching a 32.7% higher score than previous state-of-the-art methods.

3408. Sample then Identify: A General Framework for Risk Control and Assessment in Multimodal Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30685 abstract： Multimodal Large Language Models (MLLMs) exhibit promising advancements across various tasks, yet they still encounter significant trustworthiness issues. Prior studies apply Split Conformal Prediction (SCP) in language modeling to construct prediction sets with statistical guarantees. However, these methods typically rely on internal model logits or are restricted to multiple-choice settings, which hampers their generalizability and adaptability in dynamic, open-ended environments. In this paper, we introduce TRON, a two-step framework for risk control and assessment, applicable to any MLLM that supports sampling in both open-ended and closed-ended scenarios. TRON comprises two main components: (1) a novel conformal score to sample response sets of minimum size, and (2) a nonconformity score to identify high-quality responses based on self-consistency theory, controlling the error rates by two specific risk levels. Furthermore, we investigate semantic redundancy in prediction sets within open-ended contexts for the first time, leading to a promising evaluation metric for MLLMs based on average set size. Our comprehensive experiments across four Video Question-Answering (VideoQA) datasets utilizing eight MLLMs show that TRON achieves desired error rates bounded by two user-specified risk levels. Additionally, deduplicated prediction sets maintain adaptiveness while being more efficient and stable for risk assessment under different risk levels.

3409. FedLWS: Federated Learning with Adaptive Layer-wise Weight Shrinking

链接：https://iclr.cc/virtual/2025/poster/30887 abstract： In Federated Learning (FL), weighted aggregation of local models is conducted to generate a new global model, and the aggregation weights are typically normalized to 1. A recent study identifies the global weight shrinking effect in FL, indicating an enhancement in the global model’s generalization when the sum of weights (i.e., the shrinking factor) is smaller than 1, where how to learn the shrinking factor becomes crucial. However, principled approaches to this solution have not been carefully studied from the adequate consideration of privacy concerns and layer-wise distinctions. To this end, we propose a novel model aggregation strategy, Federated Learning with Adaptive Layer-wise Weight Shrinking (FedLWS), which adaptively designs the shrinking factor in a layer-wise manner and avoids optimizing the shrinking factors on a proxy dataset. We initially explored the factors affecting the shrinking factor during the training process. Then we calculate the layer-wise shrinking factors by considering the distinctions among each layer of the global model. FedLWS can be easily incorporated with various existing methods due to its flexibility. Extensive experiments under diverse scenarios demonstrate the superiority of our method over several state-of-the-art approaches, providing a promising tool for enhancing the global model in FL.

3410. DRL: Decomposed Representation Learning for Tabular Anomaly Detection

链接：https://iclr.cc/virtual/2025/poster/30521 abstract： Anomaly detection, indicating to identify the anomalies that significantly deviate from the majority normal instances of data, has been an important role in machine learning and related applications. Despite the significant success achieved in anomaly detection on image and text data, the accurate Tabular Anomaly Detection (TAD) has still been hindered due to the lack of clear prior semantic information in the tabular data. Most state-of-the-art TAD studies are along the line of reconstruction, which first reconstruct training data and then use reconstruction errors to decide anomalies; however, reconstruction on training data can still hardly distinguish anomalies due to the data entanglement in their representations. To address this problem, in this paper, we propose a novel approach Decomposed Representation Learning (DRL), to re-map data into a tailor-designed constrained space, in order to capture the underlying shared patterns of normal samples and differ anomalous patterns for TAD.Specifically, we enforce the representation of each normal sample in the latent space to be decomposed into a weighted linear combination of randomly generated orthogonal basis vectors, where these basis vectors are both data-free and training-free.Furthermore, we enhance the discriminative capability between normal and anomalous patterns in the latent space by introducing a novel constraint that amplifies the discrepancy between these two categories, supported by theoretical analysis. Finally, extensive experiments on 40 tabular datasets and 16 competing tabular anomaly detection algorithms show that our method achieves state-of-the-art performance.

3411. Gaussian-Based Instance-Adaptive Intensity Modeling for Point-Supervised Facial Expression Spotting

链接：https://iclr.cc/virtual/2025/poster/28978 abstract： Point-supervised facial expression spotting (P-FES) aims to localize facial expression instances in untrimmed videos, requiring only a single timestamp label for each instance during training. To address label sparsity, hard pseudo-labeling is often employed to propagate point labels to unlabeled frames; however, this approach can lead to confusion when distinguishing between neutral and expression frames with various intensities, which can negatively impact model performance. In this paper, we propose a two-branch framework for P-FES that incorporates a Gaussian-based instance-adaptive Intensity Modeling (GIM) module for soft pseudo-labeling. GIM models the expression intensity distribution for each instance. Specifically, we detect the pseudo-apex frame around each point label, estimate the duration, and construct a Gaussian distribution for each expression instance. We then assign soft pseudo-labels to pseudo-expression frames as intensity values based on the Gaussian distribution. Additionally, we introduce an Intensity-Aware Contrastive (IAC) loss to enhance discriminative feature learning and suppress neutral noise by contrasting neutral frames with expression frames of various intensities. Extensive experiments on the SAMM-LV and CAS(ME)^2 datasets demonstrate the effectiveness of our proposed framework. Code is available at https://github.com/KinopioIsAllIn/GIM.

3412. FlowDec: A flow-based full-band general audio codec with high perceptual quality

链接：https://iclr.cc/virtual/2025/poster/27945 abstract： We propose FlowDec, a neural full-band audio codec for general audio sampled at 48 kHz that combines non-adversarial codec training with a stochastic postfilter based on a novel conditional flow matching method. Compared to the prior work ScoreDec which is based on score matching, we generalize from speech to general audio and move from 24 kbit/s to as low as 4 kbit/s, while improving output quality and reducing the required postfilter DNN evaluations from 60 to 6 without any fine-tuning or distillation techniques. We provide theoretical insights and geometric intuitions for our approach in comparison to ScoreDec as well as another recent work that uses flow matching, and conduct ablation studies on our proposed components. We show that FlowDec is a competitive alternative to the recent GAN-dominated stream of neural codecs, achieving FAD scores better than those of the established GAN-based codec DAC and listening test scores that are on par, and producing qualitatively more natural reconstructions for speech and harmonic structures in music.

3413. MuseGNN: Forming Scalable, Convergent GNN Layers that Minimize a Sampling-Based Energy

链接：https://iclr.cc/virtual/2025/poster/30256 abstract： Among the many variants of graph neural network (GNN) architectures capable of modeling data with cross-instance relations, an important subclass involves layers designed such that the forward pass iteratively reduces a graph-regularized energy function of interest. In this way, node embeddings produced at the output layer dually serve as both predictive features for solving downstream tasks (e.g., node classification) and energy function minimizers that inherit transparent, exploitable inductive biases and interpretability. However, scaling GNN architectures constructed in this way remains challenging, in part because the convergence of the forward pass may involve models with considerable depth. To tackle this limitation, we propose a sampling-based energy function and scalable GNN layers that iteratively reduce it, guided by convergence guarantees in certain settings. We also instantiate a full GNN architecture based on these designs, and the model achieves competitive accuracy and scalability when applied to the largest publicly-available node classification benchmark exceeding 1TB in size. Our source code is available at https://github.com/haitian-jiang/MuseGNN.

3414. Fine-Tuning Discrete Diffusion Models via Reward Optimization with Applications to DNA and Protein Design

链接：https://iclr.cc/virtual/2025/poster/30308 abstract： Recent studies have demonstrated the strong empirical performance of diffusion models on discrete sequences (i.e., discrete diffusion models) across domains such as natural language and biological sequence generation. For example, in the protein inverse folding task, where the goal is to generate a protein sequence from a given backbone structure, conditional diffusion models have achieved impressive results in generating "natural" sequences that fold back into the original structure. However, practical design tasks often require not only modeling a conditional distribution but also optimizing specific task objectives. For instance, in the inverse folding task, we may prefer proteins with high stability. To address this, we consider the scenario where we have pre-trained discrete diffusion models that can generate "natural" sequences, as well as reward models that map sequences to task objectives. We then formulate the reward maximization problem within discrete diffusion models, analogous to reinforcement learning (RL), while minimizing the KL divergence against pre-trained diffusion models to preserve naturalness. To solve this RL problem, we propose a novel algorithm that enables direct backpropagation of rewards through entire trajectories generated by diffusion models, by making the originally non-differentiable trajectories differentiable using the Gumbel-Softmax trick. Our theoretical analysis indicates that our approach can generate sequences that are both "natural" (i.e., have a high probability under a pre-trained model) and yield high rewards. While similar tasks have been recently explored in diffusion models for continuous domains, our work addresses unique algorithmic and theoretical challenges specific to discrete diffusion models, which arise from their foundation in continuous-time Markov chains rather than Brownian motion. Finally, we demonstrate the effectiveness of our algorithm in generating DNA and protein sequences that optimize enhancer activity and protein stability, respectively, important tasks for gene therapies and protein-based therapeutics. The code is available at https://github.com/ChenyuWang-Monica/DRAKES.

3415. AndroidWorld: A Dynamic Benchmarking Environment for Autonomous Agents

链接：https://iclr.cc/virtual/2025/poster/28677 abstract： Autonomous agents that execute human tasks by controlling computers can enhance human productivity and application accessibility. However, progress in this field will be driven by realistic and reproducible benchmarks. We present AndroidWorld, a fully functional Android environment that provides reward signals for 116 programmatic tasks across 20 real-world Android apps. Unlike existing interactive environments, which provide a static test set, AndroidWorld dynamically constructs tasks that are parameterized and expressed in natural language in unlimited ways, thus enabling testing on a much larger and more realistic suite of tasks. To ensure reproducibility, each task includes dedicated initialization, success-checking, and tear-down logic, which modifies and inspects the device’s system state.We experiment with baseline agents to test AndroidWorld and provide initial results on the benchmark. Our best agent can complete 30.6% of AndroidWorld's tasks, leaving ample room for future work. Furthermore, we adapt a popular desktop web agent to work on Android, which we find to be less effective on mobile, suggesting future research is needed to achieve universal, cross-platform agents. Finally, we also conduct a robustness analysis, showing that task variations can significantly affect agent performance, demonstrating that without such testing, agent performance metrics may not fully reflect practical challenges. AndroidWorld and the experiments in this paper are available at https://github.com/google-research/android_world.

3416. AutoBencher: Towards Declarative Benchmark Construction

链接：https://iclr.cc/virtual/2025/poster/27700 abstract： We present AutoBencher, a declarative framework for automatic benchmark construction, and use it to scalably discover novel insights and vulnerabilities of existing language models. Concretely, given a few desiderata of benchmarks (e.g., question difficulty, topic salience), we operationalize each desideratum and cast benchmark creation as an optimization problem. Specifically, we experiment with two settings with different optimization objectives: (i) for capability evaluation, we declare the goal of finding a salient, difficult dataset that induces novel performance patterns; (ii) for safety evaluation, we declare the goal of finding a dataset of unsafe prompts that existing LMs fail to decline. To tackle this optimization problem, we use a language model to iteratively propose and refine dataset descriptions, which are then used to generate topic-specific questions and answers. These descriptions are optimized to improve the declared desiderata. We use AutoBencher (powered by GPT-4) to create datasets for math, multilinguality, knowledge, and safety. The scalability of AutoBencher allows it to test fine-grained categories and tail knowledge, creating datasets that elicit 22% more model errors (i.e., difficulty) than existing benchmarks. On the novelty ends, AutoBencher also helps identify specific gaps not captured by existing benchmarks: e.g., Gemini-Pro has knowledge gaps on Permian Extinction and Fordism while GPT-4o fails to decline harmful requests about cryptocurrency scams.

3417. Avoid Overclaims: Summary of Complexity Bounds for Algorithms in Minimization and Minimax Optimization

链接：https://iclr.cc/virtual/2025/poster/31341 abstract： In this blog, we summarize the upper and lower complexity bounds of first-order algorithms in minimization and minimax optimization problems. Within the classical oracle model framework, we review the state-of-the-art upper and lower bound results in various settings, aiming to identify gaps in existing research. With the rapid development of applications like machine learning and operation research, we further identify some recent works that revised the classical settings of optimization algorithms study.

3418. On the Identification of Temporal Causal Representation with Instantaneous Dependence

链接：https://iclr.cc/virtual/2025/poster/31125 abstract： Temporally causal representation learning aims to identify the latent causal process from time series observations, but most methods require the assumption that the latent causal processes do not have instantaneous relations. Although some recent methods achieve identifiability in the instantaneous causality case, they require either interventions on the latent variables or grouping of the observations, which are in general difficult to obtain in real-world scenarios. To fill this gap, we propose an \textbf{ID}entification framework for instantane\textbf{O}us \textbf{L}atent dynamics (\textbf{IDOL}) by imposing a sparse influence constraint that the latent causal processes have sparse time-delayed and instantaneous relations. Specifically, we establish identifiability results of the latent causal process based on sufficient variability and the sparse influence constraint by employing contextual information of time series data. Based on these theories, we incorporate a temporally variational inference architecture to estimate the latent variables and a gradient-based sparsity regularization to identify the latent causal process. Experimental results on simulation datasets illustrate that our method can identify the latent causal process. Furthermore, evaluations on multiple human motion forecasting benchmarks with instantaneous dependencies indicate the effectiveness of our method in real-world settings.

3419. Poison-splat: Computation Cost Attack on 3D Gaussian Splatting

链接：https://iclr.cc/virtual/2025/poster/30377 abstract：

3420. FaithEval: Can Your Language Model Stay Faithful to Context, Even If "The Moon is Made of Marshmallows"

链接：https://iclr.cc/virtual/2025/poster/29464 abstract： Ensuring faithfulness to context in large language models (LLMs) and retrieval-augmented generation (RAG) systems is crucial for reliable deployment in real-world applications, as incorrect or unsupported information can erode user trust. Despite advancements on standard benchmarks, faithfulness hallucination—where models generate responses misaligned with the provided context—remains a significant challenge. In this work, we introduce FaithEval, a novel and comprehensive benchmark tailored to evaluate the faithfulness of LLMs in contextual scenarios across three diverse tasks: unanswerable, inconsistent, and counterfactual contexts. These tasks simulate real-world challenges where retrieval mechanisms may surface incomplete, contradictory, or fabricated information. FaithEval comprises 4.9K high-quality problems in total, validated through a rigorous four-stage context construction and validation framework, employing both LLM-based auto-evaluation and human validation. Our extensive study across a wide range of open-source and proprietary models reveals that even state-of-the-art models often struggle to remain faithful to the given context, and that larger models do not necessarily exhibit improved faithfulness. Code is available at: https://github.com/SalesforceAIResearch/FaithEval.

3421. Robots Pre-train Robots: Manipulation-Centric Robotic Representation from Large-Scale Robot Datasets

链接：https://iclr.cc/virtual/2025/poster/27726 abstract： The pre-training of visual representations has enhanced the efficiency of robot learning. Due to the lack of large-scale in-domain robotic datasets, prior works utilize in-the-wild human videos to pre-train robotic visual representation. Despite their promising results, representations from human videos are inevitably subject to distribution shifts and lack the dynamics information crucial for task completion. We first evaluate various pre-trained representations in terms of their correlation to the downstream robotic manipulation tasks (i.e., manipulation centricity). Interestingly, we find that the "manipulation centricity" is a strong indicator of success rates when applied to downstream tasks. Drawing from these findings, we propose Manipulation Centric Representation (MCR), a foundation representation learning framework capturing both visual features and the dynamics information such as actions and proprioceptions of manipulation tasks to improve manipulation centricity. Specifically, we pre-train a visual encoder on the DROID robotic dataset and leverage motion-relevant data such as robot proprioceptive states and actions. We introduce a novel contrastive loss that aligns visual observations with the robot's proprioceptive state-action dynamics, combined with an action prediction loss and a time contrastive loss during pre-training. Empirical results across four simulation domains with 20 robotic manipulation tasks demonstrate that MCR outperforms the strongest baseline by 14.8\%. Additionally, MCR significantly boosts the success rate in three real-world manipulation tasks by 76.9\%. Project website: robots-pretrain-robots.github.io

3422. Multi-Label Node Classification with Label Influence Propagation

链接：https://iclr.cc/virtual/2025/poster/31071 abstract： Graphs are a complex and versatile data structure used across various domains, with possibly multi-label nodes playing a particularly crucial role. Examples include proteins in PPI networks with multiple functions and users in social or e-commerce networks exhibiting diverse interests. Tackling multi-label node classification (MLNC) on graphs has led to the development of various approaches. Some methods leverage graph neural networks (GNNs) to exploit label co-occurrence correlations, while others incorporate label embeddings to capture label proximity. However, these approaches fail to account for the intricate influences between labels in non-Euclidean graph data.To address this issue, we decompose the message passing process in GNNs into two operations: propagation and transformation. We then conduct a comprehensive analysis and quantification of the influence correlations between labels in each operation. Building on these insights, we propose a novel model, Label Influence Propagation (LIP). Specifically, we construct a label influence graph based on the integrated label correlations. Then, we propagate high-order influences through this graph, dynamically adjusting the learning process by amplifying labels with positive contributions and mitigating those with negative influence.Finally, our framework is evaluated on comprehensive benchmark datasets, consistently outperforming SOTA methods across various settings, demonstrating its effectiveness on MLNC tasks.

3423. StructRAG: Boosting Knowledge Intensive Reasoning of LLMs via Inference-time Hybrid Information Structurization

链接：https://iclr.cc/virtual/2025/poster/30265 abstract： Retrieval-augmented generation (RAG) is a key means to effectively enhance large language models (LLMs) in many knowledge-based tasks. However, existing RAG methods struggle with knowledge-intensive reasoning tasks, because useful information required to these tasks are badly scattered. This characteristic makes it difficult for existing RAG methods to accurately identify key information and perform global reasoning with such noisy augmentation.In this paper, motivated by the cognitive theories that humans convert raw information into various structured knowledge when tackling knowledge-intensive reasoning, we proposes a new framework, StructRAG, which can identify the optimal structure type for the task at hand, reconstruct original documents into this structured format, and infer answers based on the resulting structure. Extensive experiments across various knowledge-intensive tasks show that StructRAG achieves state-of-the-art performance, particularly excelling in challenging scenarios, demonstrating its potential as an effective solution for enhancing LLMs in complex real-world applications.

3424. Generalizing Weisfeiler-Lehman Kernels to Subgraphs

链接：https://iclr.cc/virtual/2025/poster/30220 abstract： Subgraph representation learning has been effective in solving various real-world problems. However, current graph neural networks (GNNs) produce suboptimal results for subgraph-level tasks due to their inability to capture complex interactions within and between subgraphs. To provide a more expressive and efficient alternative, we propose WLKS, a Weisfeiler-Lehman (WL) kernel generalized for subgraphs by applying the WL algorithm on induced k-hop neighborhoods. We combine kernels across different k-hop levels to capture richer structural information that is not fully encoded in existing models. Our approach can balance expressiveness and efficiency by eliminating the need for neighborhood sampling. In experiments on eight real-world and synthetic benchmarks, WLKS significantly outperforms leading approaches on five datasets while reducing training time, ranging from 0.01x to 0.25x compared to the state-of-the-art.

3425. Digi-Q: Learning VLM Q-Value Functions for Training Device-Control Agents

链接：https://iclr.cc/virtual/2025/poster/30502 abstract： While a number of existing approaches for building foundation model agents rely on prompting or fine-tuning with human demonstrations, it is not sufficient in dynamic environments (e.g., mobile device control). On-policy reinforcement learning (RL) should address these limitations, but collecting actual rollouts in an environment is often undesirable in truly open-ended agentic problems such as mobile device control or interacting with humans, where each unit of interaction is associated with a cost. In such scenarios, a method for policy learning that can utilize off-policy experience by learning a trained action-value function is much more effective. In this paper, we develop an approach, called Digi-Q, to train VLM-based action-value Q-functions which are then used to extract the agent policy. We study our approach in the mobile device control setting. Digi-Q trains the Q-function using offline temporal-difference (TD) learning, on top of frozen, intermediate-layer features of a VLM. Compared to fine-tuning the whole VLM, this approach saves us compute and enhances scalability. To make the VLM features amenable for representing the Q-function, we need to employ an initial phase of fine-tuning to amplify coverage over actionable information needed for value function. Once trained, we use this Q-function via a Best-of-N policy extraction operator that imitates the best action out of multiple candidate actions from the current policy as ranked by the value function, enabling policy improvement without environment interaction. Digi-Q outperforms several prior methods on user-scale device control tasks in Android-in-the-Wild, attaining 21.2% improvement over prior best-performing method. In some cases, our Digi-Q ap-proach already matches state-of-the-art RL methods that require interaction. The project is open-sourced at https://github.com/DigiRL-agent/digiq

3426. CtrLoRA: An Extensible and Efficient Framework for Controllable Image Generation

链接：https://iclr.cc/virtual/2025/poster/31087 abstract： Recently, large-scale diffusion models have made impressive progress in text-to-image (T2I) generation. To further equip these T2I models with fine-grained spatial control, approaches like ControlNet introduce an extra network that learns to follow a condition image. However, for every single condition type, ControlNet requires independent training on millions of data pairs with hundreds of GPU hours, which is quite expensive and makes it challenging for ordinary users to explore and develop new types of conditions. To address this problem, we propose the CtrLoRA framework, which trains a Base ControlNet to learn the common knowledge of image-to-image generation from multiple base conditions, along with condition-specific LoRAs to capture distinct characteristics of each condition. Utilizing our pretrained Base ControlNet, users can easily adapt it to new conditions, requiring as few as 1,000 data pairs and less than one hour of single-GPU training to obtain satisfactory results in most scenarios. Moreover, our CtrLoRA reduces the learnable parameters by 90% compared to ControlNet, significantly lowering the threshold to distribute and deploy the model weights. Extensive experiments on various types of conditions demonstrate the efficiency and effectiveness of our method. Codes and model weights will be released athttps://github.com/xyfJASON/ctrlora.

3427. ELFS: Label-Free Coreset Selection with Proxy Training Dynamics

链接：https://iclr.cc/virtual/2025/poster/27703 abstract： High-quality human-annotated data is crucial for modern deep learning pipelines, yet the human annotation process is both costly and time-consuming. Given a constrained human labeling budget, selecting an informative and representative data subset for labeling can significantly reduce human annotation effort. Well-performing state-of-the-art (SOTA) coreset selection methods require ground truth labels over the whole dataset, failing to reduce the human labeling burden. Meanwhile, SOTA label-free coreset selection methods deliver inferior performance due to poor geometry-based difficulty scores. In this paper, we introduce ELFS (Effective Label-Free Coreset Selection), a novel label-free coreset selection method. ELFS significantly improves label-free coreset selection by addressing two challenges: 1) ELFS utilizes deep clustering to estimate training dynamics-based data difficulty scores without ground truth labels; 2) Pseudo-labels introduce a distribution shift in the data difficulty scores, and we propose a simple but effective double-end pruning method to mitigate bias on calculated scores. We evaluate ELFS on four vision benchmarks and show that, given the same vision encoder, ELFS consistently outperforms SOTA label-free baselines. For instance, when using SwAV as the encoder, ELFS outperforms D2 by up to 10.2% in accuracy on ImageNet-1K. We make our code publicly available on GitHub.

3428. Dreamweaver: Learning Compositional World Models from Pixels

链接：https://iclr.cc/virtual/2025/poster/28949 abstract： Humans have an innate ability to decompose their perceptions of the world into objects and their attributes, such as colors, shapes, and movement patterns. This cognitive process enables us to imagine novel futures by recombining familiar concepts. However, replicating this ability in artificial intelligence systems has proven challenging, particularly when it comes to modeling videos into compositional concepts and generating unseen, recomposed futures without relying on auxiliary data, such as text, masks, or bounding boxes. In this paper, we propose Dreamweaver, a neural architecture designed to discover hierarchical and compositional representations from raw videos and generate compositional future simulations. Our approach leverages a novel Recurrent Block-Slot Unit (RBSU) to decompose videos into their constituent objects and attributes. In addition, Dreamweaver uses a multi-future-frame prediction objective to capture disentangled representations for dynamic concepts more effectively as well as static concepts. In experiments, we demonstrate our model outperforms current state-of-the-art baselines for world modeling when evaluated under the DCI framework across multiple datasets. Furthermore, we show how the modularized concept representations of our model enable compositional imagination, allowing the generation of novel videos by recombining attributes from previously seen objects. cun-bjy.github.io/dreamweaver-website

3429. On LLM Knowledge Distillation - A Comparison between Forward KL and Reverse KL

链接：https://iclr.cc/virtual/2025/poster/31335 abstract： In this blog post, we delve into knowledge distillation techniques for Large Language Models (LLMs), with a particular focus on using Kullback-Leibler (KL) Divergence as the optimization objective. Knowledge distillation is a powerful tool to reduce model size while maintaining comparable performance, making it especially useful in scenarios with constrained computational or serving resources. We specifically explore the nuances of Forward KL divergence and Reverse KL divergence, examining their roles in the distillation process. By comparing these two approaches, we aim to uncover their behaviours, strengths, and practical applications in LLM distillation.

3430. Longhorn: State Space Models are Amortized Online Learners

链接：https://iclr.cc/virtual/2025/poster/30749 abstract： The most fundamental capability of modern AI methods such as Large Language Models (LLMs) is the ability to predict the next token in a long sequence of tokens, known as “sequence modeling.” Although the Transformers model is the current dominant approach to sequence modeling, its quadratic computational cost with respect to sequence length is a significant drawback. State-space models (SSMs) offer a promising alternative due to their linear decoding efficiency and high parallelizability during training. However, existing SSMs often rely on seemingly ad hoc linear recurrence designs.In this work, we explore SSM design through the lens of online learning, conceptualizing SSMs as meta-modules for specific online learning problems. This approach links SSM design to formulating precise online learning objectives, with state transition rules derived from optimizing these objectives.Based on this insight, we introduce a novel deep SSM architecture based on the implicit update for optimizing an online regression objective. Our experimental results show that our models outperform state-of-the-art SSMs, including the Mamba model, on standard sequence modeling benchmarks and language modeling tasks.

3431. On the Hölder Stability of Multiset and Graph Neural Networks

链接：https://iclr.cc/virtual/2025/poster/29775 abstract： Extensive research efforts have been put into characterizing and constructing maximally separating multiset and graph neural networks. However, recent empirical evidence suggests the notion of separation itself doesn't capture several interesting phenomena. On the one hand, the quality of this separation may be very weak, to the extent that the embeddings of "separable" objects might even be considered identical when using fixed finite precision. On the other hand, architectures which aren't capable of separation in theory, somehow achieve separation when taking the network to be wide enough.In this work, we address both of these issues, by proposing a novel pair-wise separation quality analysis framework which is based on an adaptation of Lipschitz and Hölder stability to parametric functions. The proposed framework, which we name Hölder in expectation, allows for separation quality analysis, without restricting the analysis to embeddings that can separate all the input space simultaneously. We prove that common sum-based models are lower-Hölder in expectation, with an exponent that decays rapidly with the network's depth . Our analysis leads to adversarial examples of graphs which can be separated by three 1-WL iterations, but cannot be separated in practice by standard maximally powerful Message Passing Neural Networks (MPNNs). To remedy this, we propose two novel MPNNs with improved separation quality, one of which is lower Lipschitz in expectation. We show these MPNNs can easily classify our adversarial examples, and compare favorably with standard MPNNs on standard graph learning tasks.

3432. Uncertainty and Influence aware Reward Model Refinement for Reinforcement Learning from Human Feedback

链接：https://iclr.cc/virtual/2025/poster/28685 abstract： Reinforcement Learning from Human Feedback (RLHF) has emerged as a standard and effective approach for training large language models (LLMs) with human preferences. In this framework, a learned reward model approximates human preferences and guides policy optimization, making it crucial to develop an accurate reward model. However, without the ``true'' reward function, challenges arise when the reward model is an imperfect proxy for human preference. Since the policy optimization continuously shifts the human preference training dataset's distribution. The fixed reward model suffers from this problem of off-distribution, especially the on policy methods. While collecting new preference data can mitigate this issue, it is costly and challenging to optimize. Thus, reusing the policy interaction samples becomes a possible way to further refine the reward model. To tackle these challenges, we introduce a novel method \textbf{U}ncertainty-\textbf{G}radient based \textbf{D}ata \textbf{A}ugmentation (\textbf{UGDA} for short) to enhance reward modeling by leveraging policy samples to maintain on-distribution performance. Specifically, UGDA selects interaction samples based on the uncertainty of the reward ensembles and the gradient based influence of policy optimization. After the reward relabeling of selected samples, we use supervised learning to refine the reward ensembles, then get the retrained policy. Extensive experiments demonstrate that by leveraging UGDA to select a few samples without the costly human preference data collection, we can improve the ability of the policy and surpass the state-of-the-art methods.

3433. 3D StreetUnveiler with Semantic-aware 2DGS - a simple baseline

链接：https://iclr.cc/virtual/2025/poster/30301 abstract： Unveiling an empty street from crowded observations captured by in-car cameras is crucial for autonomous driving. However, removing all temporarily static objects, such as stopped vehicles and standing pedestrians, presents a significant challenge. Unlike object-centric 3D inpainting, which relies on thorough observation in a small scene, street scene cases involve long trajectories that differ from previous 3D inpainting tasks. The camera-centric moving environment of captured videos further complicates the task due to the limited degree and time duration of object observation. To address these obstacles, we introduce StreetUnveiler to reconstruct an empty street. StreetUnveiler learns a 3D representation of the empty street from crowded observations. Our representation is based on the hard-label semantic 2D Gaussian Splatting (2DGS) for its scalability and ability to identify Gaussians to be removed. We inpaint rendered image after removing unwanted Gaussians to provide pseudo-labels and subsequently re-optimize the 2DGS. Given its temporal continuous movement, we divide the empty street scene into observed, partial-observed, and unobserved regions, which we propose to locate through a rendered alpha map. This decomposition helps us to minimize the regions that need to be inpainted. To enhance the temporal consistency of the inpainting, we introduce a novel time-reversal framework to inpaint frames in reverse order and use later frames as references for earlier frames to fully utilize the long-trajectory observations. Our experiments conducted on the street scene dataset successfully reconstructed a 3D representation of the empty street. The mesh representation of the empty street can be extracted for further applications.

3434. PivotMesh: Generic 3D Mesh Generation via Pivot Vertices Guidance

链接：https://iclr.cc/virtual/2025/poster/29379 abstract： Generating compact and sharply detailed 3D meshes poses a significant challenge for current 3D generative models. Different from extracting dense meshes from neural representation, some recent works try to model the native mesh distribution (i.e., a set of triangles), which generates more compact results as humans crafted. However, due to the complexity and variety of mesh topology, most of these methods are typically limited to generating meshes with simple geometry. In this paper, we introduce a generic and scalable mesh generation framework PivotMesh, which makes an initial attempt to extend the native mesh generation to large-scale datasets. We employ a transformer-based autoencoder to encode meshes into discrete tokens and decode them from face level to vertex level hierarchically. Subsequently, to model the complex typology, our model first learns to generate pivot vertices as coarse mesh representation and then generate the complete mesh tokens with the same auto-regressive Transformer. This reduces the difficulty compared with directly modeling the mesh distribution and further improves the model controllability. PivotMesh demonstrates its versatility by effectively learning from both small datasets like Shapenet, and large-scale datasets like Objaverse and Objaverse-xl. Extensive experiments indicate that PivotMesh can generate compact and sharp 3D meshes across various categories, highlighting its great potential for native mesh modeling.

3435. EMOS: Embodiment-aware Heterogeneous Multi-robot Operating System with LLM Agents

链接：https://iclr.cc/virtual/2025/poster/30375 abstract： Heterogeneous multi-robot systems (HMRS) have emerged as a powerful ap-proach for tackling complex tasks that single robots cannot manage alone. Currentlarge-language-model-based multi-agent systems (LLM-based MAS) have shownsuccess in areas like software development and operating systems, but applyingthese systems to robot control presents unique challenges. In particular, the ca-pabilities of each agent in a multi-robot system are inherently tied to the physicalcomposition of the robots, rather than predefined roles. To address this issue,we introduce a novel multi-agent framework designed to enable effective collab-oration among heterogeneous robots with varying embodiments and capabilities,along with a new benchmark named Habitat-MAS. One of our key designs isRobot Resume: Instead of adopting human-designed role play, we propose a self-prompted approach, where agents comprehend robot URDF files and call robotkinematics tools to generate descriptions of their physics capabilities to guidetheir behavior in task planning and action execution. The Habitat-MAS bench-mark is designed to assess how a multi-agent framework handles tasks that requireembodiment-aware reasoning, which includes 1) manipulation, 2) perception, 3)navigation, and 4) comprehensive multi-floor object rearrangement. The experi-mental results indicate that the robot's resume and the hierarchical design of ourmulti-agent system are essential for the effective operation of the heterogeneousmulti-robot system within this intricate problem context.

3436. Scaling Stick-Breaking Attention: An Efficient Implementation and In-depth Study

链接：https://iclr.cc/virtual/2025/poster/28218 abstract： The self-attention mechanism traditionally relies on the softmax operator, necessitating positional embeddings like RoPE, or position biases to account for token order.But current methods using still face length generalisation challenges.We investigate an alternative attention mechanism based on the stick-breaking process in larger scale settings.The method works as follows: For each token before the current, we determine a break point, which represents the proportion of the stick, the weight of the attention, to allocate to the current token.We repeat this on the remaining stick, until all tokens are allocated a weight, resulting in a sequence of attention weights.This process naturally incorporates recency bias, which has linguistic motivations for grammar parsing (Shen et al., 2017).We study the implications of replacing the conventional softmax-based attention mechanism with stick-breaking attention.We then discuss implementation of numerically stable stick-breaking attention and adapt Flash Attention to accommodate this mechanism.When used as a drop-in replacement for current softmax+RoPE attention systems, we find that stick-breaking attention performs competitively with current methods on length generalisation and downstream tasks.Stick-breaking also performs well at length generalisation, allowing a model trained with $2^{11}$ context window to perform well at $2^{14}$ with perplexity improvements.

3437. Fiddler: CPU-GPU Orchestration for Fast Inference of Mixture-of-Experts Models

链接：https://iclr.cc/virtual/2025/poster/29900 abstract： Large Language Models (LLMs) with the Mixture-of-Experts (MoE) architectures have shown promising performance on various tasks. However, due to the huge model sizes, running them in resource-constrained environments where the GPU memory is not abundant is challenging. Some existing systems propose to use CPU resources to solve that, but they either suffer from the significant overhead of frequently moving data between CPU and GPU, or fail to consider distinct characteristics of CPUs and GPUs. This paper proposes Fiddler, a resource-efficient inference system for MoE models with limited GPU resources. Fiddler strategically utilizes CPU and GPU resources by determining the optimal execution strategy. Our evaluation shows that, unlike state-of-the-art systems that optimize for specific scenarios such as single batch inference or long prefill, Fiddler performs better in all scenarios. Compared against different baselines, Fiddler achieves 1.26 times speed up in single batch inference, 1.30 times in long prefill processing, and 11.57 times in beam search inference. The code of Fiddler is publicly available at https://github.com/efeslab/fiddler.

3438. Free Hunch: Denoiser Covariance Estimation for Diffusion Models Without Extra Costs

链接：https://iclr.cc/virtual/2025/poster/31019 abstract： The covariance for clean data given a noisy observation is an important quantity in many training-free guided generation methods for diffusion models. Current methods require heavy test-time computation, altering the standard diffusion training process or denoiser architecture, or making heavy approximations. We propose a new framework that sidesteps these issues by using covariance information that is available for free from training data and the curvature of the generative trajectory, which is linked to the covariance through the second-order Tweedie's formula. We integrate these sources of information using (i) a novel method to transfer covariance estimates across noise levels and (ii) low-rank updates in a given noise level. We validate the method on linear inverse problems, where it outperforms recent baselines, especially with fewer diffusion steps.

3439. Scale-Aware Contrastive Reverse Distillation for Unsupervised Medical Anomaly Detection

链接：https://iclr.cc/virtual/2025/poster/30228 abstract： Unsupervised anomaly detection using deep learning has garnered significant research attention due to its broad applicability, particularly in medical imaging where labeled anomalous data are scarce. While earlier approaches leverage generative models like autoencoders and generative adversarial networks (GANs), they often fall short due to overgeneralization. Recent methods explore various strategies, including memory banks, normalizing flows, self-supervised learning, and knowledge distillation, to enhance discrimination. Among these, knowledge distillation, particularly reverse distillation, has shown promise. Following this paradigm, we propose a novel scale-aware contrastive reverse distillation model that addresses two key limitations of existing reverse distillation methods: insufficient feature discriminability and inability to handle anomaly scale variations. Specifically, we introduce a contrastive student-teacher learning approach to derive more discriminative representations by generating and exploring out-of-normal distributions. Further, we design a scale adaptation mechanism to softly weight contrastive distillation losses at different scales to account for the scale variation issue. Extensive experiments on benchmark datasets demonstrate state-of-the-art performance, validating the efficacy of the proposed method. The code will be made publicly available.

3440. Generalized Principal-Agent Problem with a Learning Agent

链接：https://iclr.cc/virtual/2025/poster/29975 abstract： Generalized principal-agent problems, including Stackelberg games, contract design, and Bayesian persuasion, are a class of economic problems where an agent best responds to a principal's committed strategy. We study repeated generalized principal-agent problems under the assumption that the principal does not have commitment power and the agent uses algorithms to learn to respond to the principal. We reduce this problem to a one-shot generalized principal-agent problem where the agent approximately best responds. Using this reduction, we show that: (1) if the agent uses contextual no-regret learning algorithms with regret $\mathrm{Reg}(T)$, then the principal can guarantee utility at least $U^ - \Theta\big(\sqrt{\tfrac{\mathrm{Reg}(T)}{T}}\big)$, where $U^$ is the principal's optimal utility in the classic model with a best-responding agent.(2) If the agent uses contextual no-swap-regret learning algorithms with swap-regret $\mathrm{SReg}(T)$, then the principal cannot obtain utility more than $U^ + O(\frac{\mathrm{SReg(T)}}{T})$. But (3) if the agent uses mean-based learning algorithms (which can be no-regret but not no-swap-regret), then the principal can sometimes do significantly better than $U^$.These results not only refine previous results in Stackelberg games and contract design, but also lead to new results for Bayesian persuasion with a learning agent and all generalized principal-agent problems where the agent does not have private information.

3441. Innovative Thinking, Infinite Humor: Humor Research of Large Language Models through Structured Thought Leaps

链接：https://iclr.cc/virtual/2025/poster/30526 abstract： Humor is previously regarded as a gift exclusive to humans for the following reasons. Humor is a culturally nuanced aspect of human language, presenting challenges for its understanding and generation.Humor generation necessitates a multi-hop reasoning process, with each hop founded on proper rationales. Although many studies, such as those related to GPT-o1, focus on logical reasoning with reflection and correction, they still fall short in humor generation. Due to the sparsity of the knowledge graph in creative thinking, it is arduous to achieve multi-hop reasoning.Consequently, in this paper, we propose a more robust framework for addressing the humor reasoning task, named LoL. LoL aims to inject external information to mitigate the sparsity of the knowledge graph, thereby enabling multi-hop reasoning. In the first stage of LoL, we put forward an automatic instruction-evolution method to incorporate the deeper and broader thinking processes underlying humor.Judgment-oriented instructions are devised to enhance the model's judgment capability, dynamically supplementing and updating the sparse knowledge graph. Subsequently, through reinforcement learning, the reasoning logic for each online-generated response is extracted using GPT-4o. In this process, external knowledge is re-introduced to aid the model in logical reasoning and the learning of human preferences.Finally, experimental results indicate that the combination of these two processes can enhance both the model's judgment ability and its generative capacity.These findings deepen our comprehension of the creative capabilities of large language models (LLMs) and offer approaches to boost LLMs' creative abilities for cross-domain innovative applications.

3442. Near-optimal Active Regression of Single-Index Models

链接：https://iclr.cc/virtual/2025/poster/28706 abstract： The active regression problem of the single-index model is to solve $\min_x \lVert f(Ax)-b\rVert_p$, where A is fully accessible and b can only be accessed via entry queries, with the goal of minimizing the number of queries to the entries of b.When f is Lipschitz, previous results only obtain constant-factor approximations. This work presents the first algorithm that provides a $(1+\varepsilon)$-approximation solution by querying $\tilde{O}(d^{\frac{p}{2}\vee 1}/\varepsilon^{p\vee 2})$ entries of b. This query complexity is also shown to be optimal up to logarithmic factors for $p\in [1,2]$ and the ε-dependence of $1/\varepsilon^p$ is shown to be optimal for $p>2$.

3443. Few for Many: Tchebycheff Set Scalarization for Many-Objective Optimization

链接：https://iclr.cc/virtual/2025/poster/29845 abstract： Multi-objective optimization can be found in many real-world applications where some conflicting objectives can not be optimized by a single solution. Existing optimization methods often focus on finding a set of Pareto solutions with different optimal trade-offs among the objectives. However, the required number of solutions to well approximate the whole Pareto optimal set could be exponentially large with respect to the number of objectives, which makes these methods unsuitable for handling many optimization objectives. In this work, instead of finding a dense set of Pareto solutions, we propose a novel Tchebycheff set scalarization method to find a few representative solutions (e.g., 5) to cover a large number of objectives (e.g., >100) in a collaborative and complementary manner. In this way, each objective can be well addressed by at least one solution in the small solution set. In addition, we further develop a smooth Tchebycheff set scalarization approach for efficient optimization with good theoretical guarantees. Experimental studies on different problems with many optimization objectives demonstrate the effectiveness of our proposed method.

3444. Let Your Features Tell The Differences: Understanding Graph Convolution By Feature Splitting

链接：https://iclr.cc/virtual/2025/poster/30186 abstract： Graph Neural Networks (GNNs) have demonstrated strong capabilities in processing structured data. While traditional GNNs typically treat each feature dimension equally important during graph convolution, we raise an important question: **Is the graph convolution operation equally beneficial for each feature?** If not, the convolution operation on certain feature dimensions can possibly lead to harmful effects, even worse than convolution-free models. Therefore, it is required to distinguish convolution-favored and convolution-disfavored features. Traditional feature selection methods mainly focus on identifying informative features or reducing redundancy, but they are not suitable for structured data as they overlook graph structures. In graph community, some studies have investigated the performance of GNN with respect to node features using feature homophily metrics, which assess feature consistency across graph topology. Unfortunately, these metrics do not effectively align with GNN performance and cannot be reliably used for feature selection in GNNs. To address these limitations, we introduce a novel metric, Topological Feature Informativeness (TFI), to distinguish GNN-favored and GNN-disfavored features, where its effectiveness is validated through both theoretical analysis and empirical observations. Based on TFI, we propose a simple yet effective Graph Feature Selection (GFS) method, which processes GNN-favored and GNN-disfavored features with GNNs and non-GNN models separately. Compared to original GNNs, GFS significantly improves the extraction of useful topological information from each feature with comparable computational costs. Extensive experiments show that after applying GFS to $\textbf{8}$ baseline and state-of-the-art (SOTA) GNN architectures across $\textbf{10}$ datasets, $\textbf{90\%}$ of the GFS-augmented cases show significant performance boosts. Furthermore, our proposed TFI metric outperforms other feature selection methods for GFS. These results verify the effectiveness of both GFS and TFI. Additionally, we demonstrate that GFS's improvements are robust to hyperparameter tuning, highlighting its potential as a universally valid method for enhancing various GNN architectures.

3445. Multiagent Finetuning: Self Improvement with Diverse Reasoning Chains

链接：https://iclr.cc/virtual/2025/poster/30086 abstract： Large language models (LLMs) have achieved remarkable performance in recent years but are fundamentally limited by the underlying training data. To improve models beyond the training data, recent works have explored how LLMs can be used to generate synthetic data for autonomous self-improvement. However, successive steps of self-improvement can reach a point of diminishing returns. In this work, we propose a complementary approach towards self-improvement where finetuning is applied to a multiagent society of language models. A group of language models, all starting from the same base model, are independently specialized by updating each one using data generated through multiagent interactions among the models. By training each model on independent sets of data, we illustrate how this approach enables specialization across models and diversification over the set of models. As a result, our overall system is able to preserve diverse reasoning chains and autonomously improve over many more rounds of fine-tuning than single-agent self-improvement methods. We quantitatively illustrate the efficacy of the approach across a wide suite of reasoning tasks.

3446. AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark

链接：https://iclr.cc/virtual/2025/poster/28051 abstract： Video detailed captioning is a key task which aims to generate comprehensive and coherent textual descriptions of video content, benefiting both video understanding and generation. In this paper, we propose AuroraCap, a video captioner based on a large multimodal model. We follow the simplest architecture design without additional parameters for temporal modeling. To address the overhead caused by lengthy video sequences, we implement the token merging strategy, reducing the number of input visual tokens. Surprisingly, we found that this strategy results in little performance loss. AuroraCap shows superior performance on various video and image captioning benchmarks, for example, obtaining a CIDEr of 88.9 on Flickr30k, beating GPT-4V (55.3) and Gemini-1.5 Pro (82.2). However, existing video caption benchmarks only include simple descriptions, consisting of a few dozen words, which limits research in this field. Therefore, we develop VDC, a video detailed captioning benchmark with over one thousand carefully annotated structured captions. In addition, we propose a new LLM-assisted metric VDCscore for bettering evaluation, which adopts a divide-and-conquer strategy to transform long caption evaluation into multiple short question-answer pairs. With the help of human Elo ranking, our experiments show that this benchmark better correlates with human judgments of video detailed captioning quality.

3447. Planning Anything with Rigor: General-Purpose Zero-Shot Planning with LLM-based Formalized Programming

链接：https://iclr.cc/virtual/2025/poster/31260 abstract： While large language models (LLMs) have recently demonstrated strong potential in solving planning problems, there is a trade-off between flexibility and complexity. LLMs, as zero-shot planners themselves, are still not capable of directly generating valid plans for complex planning problems such as multi-constraint or long-horizon tasks. On the other hand, many frameworks aiming to solve complex planning problems often rely on task-specific preparatory efforts, such as task-specific in-context examples and pre-defined critics/verifiers, which limits their cross-task generalization capability. In this paper, we tackle these challenges by observing that the core of many planning problems lies in optimization problems: searching for the optimal solution (best plan) with goals subject to constraints (preconditions and effects of decisions). With LLMs' commonsense, reasoning, and programming capabilities, this opens up the possibilities of a universal LLM-based approach to planning problems. Inspired by this observation, we propose LLMFP, a general-purpose framework that leverages LLMs to capture key information from planning problems and formally formulate and solve them as optimization problems from scratch, with no task-specific examples needed. We apply LLMFP to 9 planning problems, ranging from multi-constraint decision making to multi-step planning problems, and demonstrate that LLMFP achieves on average 83.7\% and 86.8\% optimal rate across 9 tasks for GPT-4o and Claude 3.5 Sonnet, significantly outperforming the best baseline (direct planning with OpenAI o1-preview) with 37.6\% and 40.7\% improvements. We also validate components of LLMFP with ablation experiments and analyzed the underlying success and failure reasons.

3448. Truncated Consistency Models

链接：https://iclr.cc/virtual/2025/poster/29213 abstract： Consistency models have recently been introduced to accelerate the generation speed of diffusion models by directly predicting the solution (data) of the probability flow ODE (PF ODE) from initial noise.However, the training of consistency models requires learning to map all intermediate points along PF ODE trajectories to their corresponding endpoints. This task is much more challenging than the ultimate objective of one-step generation, which only concerns the PF ODE's noise-to-data mapping.We empirically find that this training paradigm limits the one-step generation performance of consistency models.To address this issue, we generalize consistency training to the truncated time range, which allows the model to ignore denoising tasks at earlier time steps and focus its capacity on generation.We propose a new parameterization of the consistency function and a two-stage training procedure that prevent the truncated-time training from collapsing to a trivial solution.Experiments on CIFAR-10 and ImageNet 64×64 datasets show that our method achieves better one-step and two-step FIDs than the state-of-the-art consistency models such as iCT-deep,using more than 2$\times$ smaller networks.

3449. Heavy-Tailed Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/28024 abstract： Diffusion models achieve state-of-the-art generation quality across many applications, but their ability to capture rare or extreme events in heavy-tailed distributions remains unclear. In this work, we show that traditional diffusion and flow-matching models with standard Gaussian priors fail to capture heavy-tailed behavior. We address this by repurposing the diffusion framework for heavy-tail estimation using multivariate Student-t distributions. We develop a tailored perturbation kernel and derive the denoising posterior based on the conditional Student-t distribution for the backward process. Inspired by γ-divergence for heavy-tailed distributions, we derive a training objective for heavy-tailed denoisers. The resulting framework introduces controllable tail generation using only a single scalar hyperparameter, making it easily tunable for diverse real-world distributions. As specific instantiations of our framework, we introduce t-EDM and t-Flow, extensions of existing diffusion and flow models that employ a Student-t prior. Remarkably, our approach is readily compatible with standard Gaussian diffusion models and requires only minimal code changes. Empirically, we show that our t-EDM and t-Flow outperform standard diffusion models in heavy-tail estimation on high-resolution weather datasets in which generating rare and extreme events is crucial.

3450. Energy-Based Diffusion Language Models for Text Generation

链接：https://iclr.cc/virtual/2025/poster/28133 abstract： Despite remarkable progress in autoregressive language models, alternative generative paradigms beyond left-to-right generation are still being actively explored. Discrete diffusion models, with the capacity for parallel generation, have recently emerged as a promising alternative. Unfortunately, these models still underperform the autoregressive counterparts, with the performance gap increasing when reducing the number of sampling steps. Our analysis reveals that this degradation is a consequence of an imperfect approximation used by diffusion models. In this work, we propose Energy-based Diffusion Language Model (EDLM), an energy-based model operating at the full sequence level for each diffusion step, introduced to improve the underlying approximation used by diffusion models. More specifically, we introduce an EBM in a residual form, and show that its parameters can be obtained by leveraging a pretrained autoregressive model or by finetuning a bidirectional transformer via noise contrastive estimation. We also propose an efficient generation algorithm via parallel important sampling. Comprehensive experiments on language modeling benchmarks show that our model can consistently outperform state-of-the-art diffusion models by a significant margin, and approaches autoregressive models' perplexity. We further show that, without any generation performance drop, our framework offers a 1.3x sampling speedup over existing diffusion models. Reproduced code is available at https://github.com/MinkaiXu/Energy-Diffusion-LLM.

3451. Think while You Generate: Discrete Diffusion with Planned Denoising

链接：https://iclr.cc/virtual/2025/poster/29942 abstract：

3452. LocoVR: Multiuser Indoor Locomotion Dataset in Virtual Reality

链接：https://iclr.cc/virtual/2025/poster/30668 abstract： Understanding human locomotion is crucial for AI agents such as robots, particularly in complex indoor home environments. Modeling human trajectories in these spaces requires insight into how individuals maneuver around physical obstacles and manage social navigation dynamics. These dynamics include subtle behaviors influenced by proxemics - the social use of space, such as stepping aside to allow others to pass or choosing longer routes to avoid collisions. Previous research has developed datasets of human motion in indoor scenes, but these are often limited in scale and lack the nuanced social navigation dynamics common in home environments. To address this, we present LocoVR, a dataset of 7000+ two-person trajectories captured in virtual reality from over 130 different indoor home environments. LocoVR provides accurate trajectory and precise spatial information, along with rich examples of socially-motivated movement behaviors. For example, the dataset captures instances of individuals navigating around each other in narrow spaces, adjusting paths to respect personal boundaries in living areas, and coordinating movements in high-traffic zones like entryways and kitchens. Our evaluation shows that LocoVR significantly enhances model performance in three practical indoor tasks utilizing human trajectories, and demonstrates predicting socially-aware navigation patterns in home environments.

3453. Quantized Spike-driven Transformer

链接：https://iclr.cc/virtual/2025/poster/30954 abstract： Spiking neural networks (SNNs) are emerging as a promising energy-efficient alternative to traditional artificial neural networks (ANNs) due to their spike-driven paradigm.However, recent research in the SNN domain has mainly focused on enhancing accuracy by designing large-scale Transformer structures, which typically rely on substantial computational resources, limiting their deployment on resource-constrained devices.To overcome this challenge, we propose a quantized spike-driven Transformer baseline (QSD-Transformer), which achieves reduced resource demands by utilizing a low bit-width parameter. Regrettably, the QSD-Transformer often suffers from severe performance degradation.In this paper, we first conduct empirical analysis and find that the bimodal distribution of quantized spike-driven self-attention (Q-SDSA) leads to spike information distortion (SID) during quantization, causing significant performance degradation. To mitigate this issue, we take inspiration from mutual information entropy and propose a bi-level optimization strategy to rectify the information distribution in Q-SDSA.Specifically, at the lower level, we introduce an information-enhanced LIF to rectify the information distribution in Q-SDSA.At the upper level, we propose a fine-grained distillation scheme for the QSD-Transformer to align the distribution in Q-SDSA with that in the counterpart ANN.By integrating the bi-level optimization strategy, the QSD-Transformer can attain enhanced energy efficiency without sacrificing its high-performance advantage.We validate the QSD-Transformer on various visual tasks, and experimental results indicate that our method achieves state-of-the-art results in the SNN domain.For instance, when compared to the prior SNN benchmark on ImageNet, the QSD-Transformer achieves 80.3\% top-1 accuracy, accompanied by significant reductions of 6.0$\times$ and 8.1$\times$ in power consumption and model size, respectively. Code is available at https://github.com/bollossom/QSD-Transformer.

3454. Can Watermarked LLMs be Identified by Users via Crafted Prompts?

链接：https://iclr.cc/virtual/2025/poster/27962 abstract： Text watermarking for Large Language Models (LLMs) has made significant progress in detecting LLM outputs and preventing misuse. Current watermarking techniques offer high detectability, minimal impact on text quality, and robustness to text editing. However, current researches lack investigation into the imperceptibility of watermarking techniques in LLM services. This is crucial as LLM providers may not want to disclose the presence of watermarks in real-world scenarios, as it could reduce user willingness to use the service and make watermarks more vulnerable to attacks. This work is the first to investigate the imperceptibility of watermarked LLMs. We design an identification algorithm called Water-Probe that detects watermarks through well-designed prompts to the LLM. Our key motivation is that current watermarked LLMs expose consistent biases under the same watermark key, resulting in similar differences across prompts under different watermark keys. Experiments show that almost all mainstream watermarking algorithms are easily identified with our well-designed prompts, while Water-Probe demonstrates a minimal false positive rate for non-watermarked LLMs. Finally, we propose that the key to enhancing the imperceptibility of watermarked LLMs is to increase the randomness of watermark key selection. Based on this, we introduce the Water-Bag strategy, which significantly improves watermark imperceptibility by merging multiple watermark keys.

3455. Persistent Pre-training Poisoning of LLMs

链接：https://iclr.cc/virtual/2025/poster/28909 abstract： Large language models are pre-trained on uncurated text datasets consisting of trillions of tokens scraped from the Web.Prior work has shown that: (1) web-scraped pre-training datasets can be practically poisoned by malicious actors; and (2) adversaries can compromise language models after poisoning fine-tuning datasets.Our work evaluates for the first time whether language models can also be \emph{compromised during pre-training}, with a focus on the persistence of pre-training attacks after models are fine-tuned as helpful and harmless chatbots (i.e., after SFT and DPO).We pre-train a series of LLMs from scratch to measure the impact of a potential poisoning adversary under four different attack objectives (denial-of-service, belief manipulation, jailbreaking, and prompt stealing), and across a wide range of model sizes (from 600M to 7B).Our main result is that poisoning only 0.1% of a model's pre-training dataset is sufficient for three out of four attacks to measurably persist through post-training. Moreover, simple attacks like denial-of-service persist through post-training with a poisoning rate of only 0.001%.

3456. Backtracking Improves Generation Safety

链接：https://iclr.cc/virtual/2025/poster/30547 abstract： Text generation has a fundamental limitation almost by definition: there is no taking back tokens that have been generated, even when they are clearly problematic.In the context of language model safety, when a partial unsafe generation is produced, language models by their nature tend to happily keep on generating similarly unsafe additional text.This is in fact how safety alignment of frontier models gets circumvented in the wild, despite great efforts in improving their safety.Deviating from the paradigm of approaching safety alignment as prevention (decreasing the probability of harmful responses), we propose backtracking, a technique that allows language models to "undo" and recover from their own unsafe generation through the introduction of a special [RESET] token.Our method can be incorporated into either SFT or DPO training to optimize helpfulness and harmlessness.We show that models trained to backtrack are consistently safer than baseline models: backtracking Llama-3-8B is four times more safe than the baseline model (6.1\% $\to$ 1.5\%) in our evaluations without regression in helpfulness.Our method additionally provides protection against four adversarial attacks including an adaptive attack, despite not being trained to do so.

3457. Human-Aligned Chess With a Bit of Search

链接：https://iclr.cc/virtual/2025/poster/29104 abstract： Chess has long been a testbed for AI's quest to match human intelligence, and in recent years, chess AI systems have surpassed the strongest humans at the game.However, these systems are not human-aligned; they are unable to match the skill levels of all human partners or model human-like behaviors beyond piece movement.In this paper, we introduce Allie, a chess-playing AI designed to bridge the gap between artificial and human intelligence in this classic game.Allie is trained on log sequences of real chess games to model the behaviors of human chess players across the skill spectrum, including non-move behaviors such as pondering times and resignationsIn offline evaluations, we find that Allie exhibits humanlike behavior: it outperforms the existing state-of-the-art in human chess move prediction and ``ponders'' at critical positions.The model learns to reliably assign reward at each game state, which can be used at inference as a reward function in a novel time-adaptive Monte-Carlo tree search (MCTS) procedure, where the amount of search depends on how long humans would think in the same positions.Adaptive search enables remarkable skill calibration; in a large-scale online evaluation against players with ratings from 1000 to 2600 Elo, our adaptive search method leads to a skill gap of only 49 Elo on average, substantially outperforming search-free and standard MCTS baselines.Against grandmaster-level (2500 Elo) opponents, Allie with adaptive search exhibits the strength of a fellow grandmaster, all while learning exclusively from humans.

3458. What Are Good Positional Encodings for Directed Graphs?

链接：https://iclr.cc/virtual/2025/poster/28146 abstract： Positional encodings (PEs) are essential for building powerful and expressive graph neural networks and graph transformers, as they effectively capture the relative spatial relationships between nodes. Although extensive research has been devoted to PEs in undirected graphs, PEs for directed graphs remain relatively unexplored. This work seeks to address this gap. We first introduce the notion of Walk Profile, a generalization of walk-counting sequences for directed graphs. A walk profile encompasses numerous structural features crucial for directed graph-relevant applications, such as program analysis and circuit performance prediction. We identify the limitations of existing PE methods in representing walk profiles and propose a novel Multi-q Magnetic Laplacian PE, which extends the Magnetic Laplacian eigenvector-based PE by incorporating multiple potential factors. The new PE can provably express walk profiles. Furthermore, we generalize prior basis-invariant neural networks to enable the stable use of the new PE in the complex domain. Our numerical experiments validate the expressiveness of the proposed PEs and demonstrate their effectiveness in solving sorting network satisfiability and performing well on general circuit benchmarks. Our code is available at https://github.com/Graph-COM/Multi-q-Maglap.

3459. Consistent Flow Distillation for Text-to-3D Generation

链接：https://iclr.cc/virtual/2025/poster/30650 abstract： Score Distillation Sampling (SDS) has made significant strides in distilling image-generative models for 3D generation. However, its maximum-likelihood-seeking behavior often leads to degraded visual quality and diversity, limiting its effectiveness in 3D applications. In this work, we propose Consistent Flow Distillation (CFD), which addresses these limitations. We begin by leveraging the gradient of the diffusion ODE or SDE sampling process to guide the 3D generation. From the gradient-based sampling perspective, we find that the consistency of 2D image flows across different viewpoints is important for high-quality 3D generation. To achieve this, we introduce multi-view consistent Gaussian noise on the 3D object, which can be rendered from various viewpoints to compute the flow gradient. Our experiments demonstrate that CFD, through consistent flows, significantly outperforms previous methods in text-to-3D generation.

3460. 3D-AffordanceLLM: Harnessing Large Language Models for Open-Vocabulary Affordance Detection in 3D Worlds

链接：https://iclr.cc/virtual/2025/poster/30275 abstract： 3D Affordance detection is a challenging problem with broad applications on various robotic tasks. Existing methods typically formulate the detection paradigm as a label-based semantic segmentation task.This paradigm relies on predefined labels and lacks the ability to comprehend complex natural language, resulting in limited generalization in open-world scene.To address these limitations, we reformulate the traditional affordance detection paradigm into \textit{Instruction Reasoning Affordance Segmentation} (IRAS) task. This task is designed to output a affordance mask region given a query reasoning text, which avoids fixed categories of input labels.We accordingly propose the \textit{3D-AffordanceLLM} (3D-ADLLM), a framework designed for reasoning affordance detection in 3D open-scene.Specifically, 3D-ADLLM introduces large language models (LLMs) to 3D affordance perception with a custom-designed decoder for generating affordance masks, thus achieving open-world reasoning affordance detection.In addition, given the scarcity of 3D affordance datasets for training large models, we seek to extract knowledge from general segmentation data and transfer it to affordance detection.Thus, we propose a multi-stage training strategy that begins with a novel pre-training task, i.e., \textit{Referring Object Part Segmentation}~(ROPS).This stage is designed to equip the model with general recognition and segmentation capabilities at the object-part level.Then followed by fine-tuning with the IRAS task, 3D-ADLLM obtains the reasoning ability for affordance detection. In summary, 3D-ADLLM leverages the rich world knowledge and human-object interaction reasoning ability of LLMs, achieving approximately an 8\% improvement in mIoU on open-vocabulary affordance detection tasks.

3461. From Exploration to Mastery: Enabling LLMs to Master Tools via Self-Driven Interactions

链接：https://iclr.cc/virtual/2025/poster/29696 abstract： Tool learning enables Large Language Models (LLMs) to interact with external environments by invoking tools, serving as an effective strategy to mitigate the limitations inherent in their pre-training data. In this process, tool documentation plays a crucial role by providing usage instructions for LLMs, thereby facilitating effective tool utilization. This paper concentrates on the critical challenge of bridging the comprehension gap between LLMs and external tools due to the inadequacies and inaccuracies inherent in existing human-centric tool documentation. We propose a novel framework, DRAFT, aimed at Dynamically Refining tool documentation through the Analysis of Feedback and Trials emanating from LLMs' interactions with external tools. This methodology pivots on an innovative trial-and-error approach, consisting of three distinct learning phases: experience gathering, learning from experience, and documentation rewriting, to iteratively enhance the tool documentation. This process is further optimized by implementing a diversity-promoting exploration strategy to ensure explorative diversity and a tool-adaptive termination mechanism to prevent overfitting while enhancing efficiency. Extensive experiments on multiple datasets demonstrate that DRAFT's iterative, feedback-based refinement significantly ameliorates documentation quality, fostering a deeper comprehension and more effective utilization of tools by LLMs. Notably, our analysis reveals that the tool documentation refined via our approach demonstrates robust cross-model generalization capabilities.

3462. Step-by-Step Reasoning for Math Problems via Twisted Sequential Monte Carlo

链接：https://iclr.cc/virtual/2025/poster/29208 abstract： Augmenting the multi-step reasoning abilities of Large Language Models (LLMs) has been a persistent challenge. Recently, verification has shown promise in improving solution consistency by evaluating generated outputs. However, current verification approaches suffer from sampling inefficiencies, requiring a large number of samples to achieve satisfactory performance. Additionally, training an effective verifier often depends on extensive process supervision, which is costly to acquire. In this paper, we address these limitations by introducing a novel verification method based on Twisted Sequential Monte Carlo (TSMC). TSMC sequentially refines its sampling effort to focus exploration on promising candidates, resulting in more efficient generation of high-quality solutions. We apply TSMC to LLMs by estimating the expected future rewards at partial solutions. This approach results in a more straightforward training target that eliminates the need for step-wise human annotations. We empirically demonstrate the advantages of our method across multiple math benchmarks, and also validate our theoretical analysis of both our approach and existing verification methods.

3463. Zero-Shot Whole-Body Humanoid Control via Behavioral Foundation Models

链接：https://iclr.cc/virtual/2025/poster/30661 abstract： Unsupervised reinforcement learning (RL) aims at pre-training models that can solve a wide range of downstream tasks in complex environments. Despite recent advancements, existing approaches suffer from several limitations: they may require running an RL process on each task to achieve a satisfactory performance, they may need access to datasets with good coverage or well-curated task-specific samples, or they may pre-train policies with unsupervised losses that are poorly correlated with the downstream tasks of interest. In this paper, we introduce FB-CPR, which regularizes unsupervised zero-shot RL based on the forward-backward (FB) method towards imitating trajectories from unlabeled behaviors. The resulting models learn useful policies imitating the behaviors in the dataset, while retaining zero-shot generalization capabilities. We demonstrate the effectiveness of FB-CPR in a challenging humanoid control problem. Training FB-CPR online with observation-only motion capture datasets, we obtain the first humanoid behavioral foundation model that can be prompted to solve a variety of whole-body tasks, including motion tracking, goal reaching, and reward optimization. The resulting model is capable of expressing human-like behaviors and it achieves competitive performance with task-specific methods while outperforming state-of-the-art unsupervised RL and model-based baselines.

3464. DisEnvisioner: Disentangled and Enriched Visual Prompt for Customized Image Generation

链接：https://iclr.cc/virtual/2025/poster/27912 abstract： In the realm of image generation, creating customized images from visual prompt with additional textual instruction emerges as a promising endeavor. However, existing methods, both tuning-based and tuning-free, struggle with interpreting the subject-essential attributes from the visual prompt. This leads to subject-irrelevant attributes infiltrating the generation process, ultimately compromising the personalization quality in both editability and ID preservation. In this paper, we present $\textbf{DisEnvisioner}$, a novel approach for effectively extracting and enriching the subject-essential features while filtering out -irrelevant information, enabling exceptional customization performance, in a $\textbf{tuning-free}$ manner and using only $\textbf{a single image}$. Specifically, the feature of the subject and other irrelevant components are effectively separated into distinctive visual tokens, enabling a much more accurate customization. Aiming to further improving the ID consistency, we enrich the disentangled features, sculpting them into a more granular representation. Experiments demonstrate the superiority of our approach over existing methods in instruction response (editability), ID consistency, inference speed, and the overall image quality, highlighting the effectiveness and efficiency of DisEnvisioner.

3465. Looped Transformers for Length Generalization

链接：https://iclr.cc/virtual/2025/poster/31126 abstract： Recent work has shown that Transformers trained from scratch can successfully solve various arithmetic and algorithmic tasks, such as adding numbers and computing parity. While these Transformers generalize well on unseen inputs of the same length, they struggle with length generalization, i.e., handling inputs of unseen lengths. In this work, we demonstrate that looped Transformers with an adaptive number of steps significantly improve length generalization. We focus on tasks with a known iterative solution, involving multiple iterations of a RASP-L operation—a length-generalizable operation that can be expressed by a finite-sized Transformer. We train looped Transformers using our proposed learning algorithm and observe that they learn highly length-generalizable solutions for various tasks.

3466. A Theoretical Perspective: How to Prevent Model Collapse in Self-consuming Training Loops

链接：https://iclr.cc/virtual/2025/poster/29338 abstract： High-quality data is essential for training large generative models, yet the vast reservoir of real data available online has become nearly depleted. Consequently, models increasingly generate their own data for further training, forming Self-consuming Training Loops (STLs). However, the empirical results have been strikingly inconsistent: some models degrade or even collapse, while others successfully avoid these failures, leaving a significant gap in theoretical understanding to explain this discrepancy. This paper introduces the intriguing notion of recursive stability and presents the first theoretical generalization analysis, revealing how both model architecture and the proportion between real and synthetic data influence the success of STLs. We further extend this analysis to transformers in in-context learning, showing that even a constant-sized proportion of real data ensures convergence, while also providing insights into optimal synthetic data sizing.

3467. Efficient Reward Poisoning Attacks on Online Deep Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/31512 abstract： We study reward poisoning attacks on online deep reinforcement learning (DRL), where the attacker is oblivious to the learning algorithm used by the agent and the dynamics of the environment. We demonstrate the intrinsic vulnerability of state-of-the-art DRL algorithms by designing a general, black-box reward poisoning framework called adversarial MDP attacks. We instantiate our framework to construct two new attacks which only corrupt the rewards for a small fraction of the total training timesteps and make the agent learn a low-performing policy. We provide a theoretical analysis of the efficiency of our attack and perform an extensive empirical evaluation. Our results show that our attacks efficiently poison agents learning in several popular classical control and MuJoCo environments with a variety of state-of-the-art DRL algorithms, such as DQN, PPO, SAC, etc.

3468. Language Models Need Inductive Biases to Count Inductively

链接：https://iclr.cc/virtual/2025/poster/28148 abstract： Counting constitutes a core skill underlying a wide range of tasks, such as formal language recognition, multi-hop reasoning and simulating algorithms. Generaliz- ing counting inductively is central to task success on out-of-distribution (OOD) instances where testing inputs are longer than those seen in training. While there is a large body of literature reporting poor length generalization in language models, few papers have tried to distill the “reasoning” failure to the simplest case of count- ing failure. We aim to provide a broader picture on whether various language model architectures can a) learn to count, and b) generalize counting inductively. This work provides extensive empirical results on architectures ranging from RNNs, Transformers, State-Space Models and RWKV. We present carefully-designed task formats, auxiliary tasks and positional embeddings to avoid limitations in general- ization with OOD-position and OOD-vocabulary. We find that while traditional RNNs trivially achieve inductive counting, Transformers have to rely on positional embeddings (PEs) to count OOD. Further analyses on interpreting the learned solution reveal that different PEs encode different inductive biases that facilitate counting in different task formats. As counting is the basis for many arguments concerning the expressivity of Transformers, our finding calls for the community to reexamine the application scope of primitive functions defined in formal charac- terizations. Finally, modern RNNs also largely underperform traditional RNNs in generalizing counting inductively, hinting at the tradeoff modern RNNs struggle to balance between parallelized training and maintaining their recurrent nature.

3469. CLDyB: Towards Dynamic Benchmarking for Continual Learning with Pre-trained Models

链接：https://iclr.cc/virtual/2025/poster/29639 abstract： The emergence of the foundation model era has sparked immense research interest in utilizing pre-trained representations for continual learning~(CL), yielding a series of strong CL methods with outstanding performance on standard evaluation benchmarks. Nonetheless, there are growing concerns regarding potential data contamination within the massive pre-training datasets. Furthermore, the static nature of standard evaluation benchmarks tends to oversimplify the complexities encountered in real-world CL scenarios, putting CL methods at risk of overfitting to these benchmarks while still lacking robustness needed for more demanding real-world applications. To solve these problems, this paper proposes a general framework to evaluate methods for Continual Learning on Dynamic Benchmarks (CLDyB). CLDyB continuously identifies inherently challenging tasks for the specified CL methods and evolving backbones, and dynamically determines the sequential order of tasks at each time step in CL using a tree-search algorithm, guided by an overarching goal to generate highly challenging task sequences for evaluation. To highlight the significance of dynamic evaluation on the CLDyB, we first simultaneously evaluate multiple state-of-the-art CL methods under CLDyB, resulting in a set of commonly challenging task sequences where existing CL methods tend to underperform. We intend to publicly release these task sequences for the CL community to facilitate the training and evaluation of more robust CL algorithms. Additionally, we perform individual evaluations of the CL methods under CLDyB, yielding informative evaluation results that reveal the specific strengths and weaknesses of each method.

3470. Unlocking the Power of Function Vectors for Characterizing and Mitigating Catastrophic Forgetting in Continual Instruction Tuning

链接：https://iclr.cc/virtual/2025/poster/28810 abstract： Catastrophic forgetting (CF) poses a significant challenge in machine learning, where a model forgets previously learned information upon learning new tasks. Despite the advanced capabilities of Large Language Models (LLMs), they continue to face challenges with CF during continual learning. The majority of existing research focuses on analyzing forgetting patterns through a singular training sequence, thereby overlooking the intricate effects that diverse tasks have on model behavior.Our study explores CF across various settings, discovering that model forgetting is influenced by both the specific training tasks and the models themselves. To this end, we interpret forgetting by examining the function vector (FV), a compact representation of functions in LLMs, offering a model-dependent indicator for the occurrence of CF. Through theoretical and empirical analyses, we demonstrated that CF in LLMs primarily stems from biases in function activation rather than the overwriting of task processing functions.Leveraging these insights, we propose a novel function vector guided training methodology, incorporating a regularization technique to stabilize the FV and mitigate forgetting. Empirical tests on four benchmarks confirm the effectiveness of our proposed training method, substantiating our theoretical framework concerning CF and model function dynamics.

3471. SD-LoRA: Scalable Decoupled Low-Rank Adaptation for Class Incremental Learning

链接：https://iclr.cc/virtual/2025/poster/30945 abstract： Continual Learning (CL) with foundation models has recently emerged as a promising paradigm to exploit abundant knowledge acquired during pre-training for tackling sequential tasks. However, existing prompt-based and Low-Rank Adaptation-based (LoRA-based) methods often require expanding a prompt/LoRA pool or retaining samples of previous tasks, which poses significant scalability challenges as the number of tasks grows. To address these limitations, we propose Scalable Decoupled LoRA (SD-LoRA) for class incremental learning, which continually separates the learning of the magnitude and direction of LoRA components without rehearsal. Our empirical and theoretical analysis reveals that SD-LoRA tends to follow a low-loss trajectory and converges to an overlapping low-loss region for all learned tasks, resulting in an excellent stability-plasticity trade-off. Building upon these insights, we introduce two variants of SD-LoRA with further improved parameter efficiency. All parameters of SD-LoRAs can be end-to-end optimized for CL objectives. Meanwhile, they support efficient inference by allowing direct evaluation with the finally trained model, obviating the need for component selection. Extensive experiments across multiple CL benchmarks and foundation models consistently validate the effectiveness of SD-LoRA. The code is available at https://github.com/WuYichen-97/SD-Lora-CL.

3472. VideoWebArena: Evaluating Long Context Multimodal Agents with Video Understanding Web Tasks

链接：https://iclr.cc/virtual/2025/poster/27960 abstract： Videos are often used to learn or extract the necessary information to completetasks in ways different than what text or static imagery can provide. However, manyexisting agent benchmarks neglect long-context video understanding, instead focus-ing on text or static image inputs. To bridge this gap, we introduce VideoWebArena(VideoWA), a benchmark for evaluating the capabilities of long-context multimodalagents for video understanding. VideoWA consists of 2,021 web agent tasks basedon manually crafted video tutorials, which total almost four hours of content. Forour benchmark, we define a taxonomy of long-context video-based agent tasks withtwo main areas of focus: skill retention and factual retention. While skill retentiontasks evaluate whether an agent can use a given human demonstration to completea task efficiently, the factual retention task evaluates whether an agent can retrieveinstruction-relevant information from a video to complete a task. We find that thebest model achieves a 13.3% success rate on factual retention tasks and 45.8% onfactual retention QA pairs—far below human success rates of 73.9% and 79.3%,respectively. On skill retention tasks, long-context models perform worse withtutorials than without, exhibiting a 5% performance decrease in WebArena tasksand a 10.3% decrease in VisualWebArena tasks. Our work highlights performancegaps in the agentic abilities of long-context multimodal models and provides as atestbed for the future development of long-context video agents.

3473. Learning-Guided Rolling Horizon Optimization for Long-Horizon Flexible Job-Shop Scheduling

链接：https://iclr.cc/virtual/2025/poster/30617 abstract： Long-horizon combinatorial optimization problems (COPs), such as the Flexible Job-Shop Scheduling Problem (FJSP), often involve complex, interdependent decisions over extended time frames, posing significant challenges for existing solvers. While Rolling Horizon Optimization (RHO) addresses this by decomposing problems into overlapping shorter-horizon subproblems, such overlap often involves redundant computations. In this paper, we present L-RHO, the first learning-guided RHO framework for COPs. L-RHO employs a neural network to intelligently fix variables that in hindsight did not need to be re-optimized, resulting in smaller and thus easier-to-solve subproblems. For FJSP, this means identifying operations with unchanged machine assignments between consecutive subproblems. Applied to FJSP, L-RHO accelerates RHO by up to 54\% while significantly improving solution quality, outperforming other heuristic and learning-based baselines. We also provide in-depth discussions and verify the desirable adaptability and generalization of L-RHO across numerous FJSP variates, distributions, online scenarios and benchmark instances. Moreover, we provide a theoretical analysis to elucidate the conditions under which learning is beneficial.

3474. GravMAD: Grounded Spatial Value Maps Guided Action Diffusion for Generalized 3D Manipulation

链接：https://iclr.cc/virtual/2025/poster/28249 abstract： Robots' ability to follow language instructions and execute diverse 3D manipulation tasks is vital in robot learning. Traditional imitation learning-based methods perform well on seen tasks but struggle with novel, unseen ones due to variability. Recent approaches leverage large foundation models to assist in understanding novel tasks, thereby mitigating this issue. However, these methods lack a task-specific learning process, which is essential for an accurate understanding of 3D environments, often leading to execution failures. In this paper, we introduce GravMAD, a sub-goal-driven, language-conditioned action diffusion framework that combines the strengths of imitation learning and foundation models. Our approach breaks tasks into sub-goals based on language instructions, allowing auxiliary guidance during both training and inference. During training, we introduce Sub-goal Keypose Discovery to identify key sub-goals from demonstrations. Inference differs from training, as there are no demonstrations available, so we use pre-trained foundation models to bridge the gap and identify sub-goals for the current task. In both phases, GravMaps are generated from sub-goals, providing GravMAD with more flexible 3D spatial guidance compared to fixed 3D positions. Empirical evaluations on RLBench show that GravMAD significantly outperforms state-of-the-art methods, with a 28.63\% improvement on novel tasks and a 13.36\% gain on tasks encountered during training. Evaluations on real-world robotic tasks further show that GravMAD can reason about real-world tasks, associate them with relevant visual information, and generalize to novel tasks. These results demonstrate GravMAD's strong multi-task learning and generalization in 3D manipulation. Video demonstrations are available at: https://gravmad.github.io.

3475. MADGEN: Mass-Spec attends to De Novo Molecular generation

链接：https://iclr.cc/virtual/2025/poster/30846 abstract： The annotation (assigning structural chemical identities) of MS/MS spectra remains a significant challenge due to the enormous molecular diversity in biological samples and the limited scope of reference databases. Currently, the vast majority of spectral measurements remain in the "dark chemical space" without structural annotations. To improve annotation, we propose MADGEN (Mass-spec Attends to De Novo Molecular GENeration), a scaffold-based method for de novo molecular structure generation guided by mass spectrometry data. MADGEN operates in two stages: scaffold retrieval and spectra-conditioned molecular generation starting with the scaffold. In the first stage, given an MS/MS spectrum, we formulate scaffold retrieval as a ranking problem and employ contrastive learning to align mass spectra with candidate molecular scaffolds. In the second stage, starting from the retrieved scaffold, we employ the MS/MS spectrum to guide an attention-based generative model to generate the final molecule. Our approach constrains the molecular generation search space, reducing its complexity and improving generation accuracy. We evaluate MADGEN on three datasets (NIST23, CANOPUS, and MassSpecGym) and evaluate MADGEN's performance with a predictive scaffold retriever and with an oracle retriever. We demonstrate the effectiveness of using attention to integrate spectral information throughout the generation process to achieve strong results with the oracle retriever.

3476. Inference-Aware Fine-Tuning for Best-of-N Sampling in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30849 abstract： Recent studies indicate that effectively utilizing inference-time compute is crucial for attaining good performance from large language models (LLMs). Specifically, the Best-of-N (BoN) inference strategy, where an LLM generates multiple responses and a verifier selects the best, has shown strong empirical performance.

Motivated by this, we develop a novel inference-aware fine-tuning paradigm, which encompasses the BoN-aware inference framework as a special case. We devise the first imitation learning and reinforcement learning (RL) methods for fine-tuning LLMs using BoN, overcoming the challenging, non-differentiable argmax operator in BoN. We empirically demonstrate that our BoN-aware models implicitly learn a per-example "meta-strategy", which interleaves best responses with more diverse responses that might be better suited to a test-time input—a process reminiscent of the exploration-exploitation trade-off in RL. Our experiments demonstrate the effectiveness of BoN-aware fine-tuning in terms of improved performance and inference-time compute. In particular, we show that our methods improve the BoN performance of Gemma 2B on Hendrycks MATH from 26.8% to 30.8%, and Pass@K from 60% to 67%.

3477. MELODI: Exploring Memory Compression for Long Contexts

链接：https://iclr.cc/virtual/2025/poster/29502 abstract： We present MELODI, a novel memory architecture designed to efficiently process long documents using short context windows. The key principle behind MELODI is to represent short-term and long-term memory as a hierarchical compression scheme across both transformer layers and context windows. Specifically, the short-term memory is achieved through recurrent compression of context windows across multiple layers, ensuring smooth transitions between windows. In contrast, the long-term memory performs further compression within a single middle layer and aggregates information across context windows, effectively consolidating crucial information from the entire history. Compared to a strong baseline - the Memorizing Transformer employing dense attention over a large long-term memory (64K key-value pairs) - our method demonstrates superior performance on various long-context datasets while remarkably reducing the memory footprint by a factor of 8.

3478. How Discrete and Continuous Diffusion Meet: Comprehensive Analysis of Discrete Diffusion Models via a Stochastic Integral Framework

链接：https://iclr.cc/virtual/2025/poster/30876 abstract： Discrete diffusion models have gained increasing attention for their ability to model complex distributions with tractable sampling and inference. However, the error analysis for discrete diffusion models remains less well-understood. In this work, we propose a comprehensive framework for the error analysis of discrete diffusion models based on Lévy-type stochastic integrals. By generalizing the Poisson random measure to that with a time-independent and state-dependent intensity, we rigorously establish a stochastic integral formulation of discrete diffusion models and provide the corresponding change of measure theorems that are intriguingly analogous to Itô integrals and Girsanov's theorem for their continuous counterparts. Our framework unifies and strengthens the current theoretical results on discrete diffusion models and obtains the first error bound for the τ-leaping scheme in KL divergence. With error sources clearly identified, our analysis gives new insight into the mathematical properties of discrete diffusion models and offers guidance for the design of efficient and accurate algorithms for real-world discrete diffusion model applications.

3479. Better autoregressive regression with LLMs via regression-aware fine-tuning

链接：https://iclr.cc/virtual/2025/poster/27796 abstract： Decoder-based large language models (LLMs) have proven highly versatile, with remarkable successes even on problems ostensibly removed from traditional language generation. One such example is solving regression problems, where the targets are real numbers rather than textual tokens. A common approach to use LLMs on such problems is to perform fine-tuning based on the cross-entropy loss, and use autoregressive sampling at inference time. Another approach relies on fine-tuning a separate predictive head with a suitable loss such as squared error. While each approach has had success, there has been limited study on principled ways of using decoder LLMs for regression. In this work, we compare different prior works under a unified view, and introduce regression-aware fine-tuning(RAFT), a novel approach based on the Bayes-optimal decision rule. We demonstrate how RAFT improves over established baselines on several benchmarks and model families.

3480. A Simple Approach to Unifying Diffusion-based Conditional Generation

链接：https://iclr.cc/virtual/2025/poster/28070 abstract： Recent progress in image generation has sparked research into controlling these models through condition signals, with various methods addressing specific challenges in conditional generation. Instead of proposing another specialized technique, we introduce a simple, unified framework to handle diverse conditional generation tasks involving a specific image-condition correlation. By learning a joint distribution over a correlated image pair (e.g. image and depth) with a diffusion model, our approach enables versatile capabilities via different inference-time sampling schemes, including controllable image generation (e.g. depth to image), estimation (e.g. image to depth), signal guidance, joint generation (image \& depth), and coarse control. Previous attempts at unification often introduce complexity through multi-stage training, architectural modification, or increased parameter counts. In contrast, our simplified formulation requires a single, computationally efficient training stage, maintains the standard model input, and adds minimal learned parameters (15% of the base model). Moreover, our model supports additional capabilities like non-spatially aligned and coarse conditioning. Extensive results show that our single model can produce comparable results with specialized methods and better results than prior unified methods. We also demonstrate that multiple models can be effectively combined for multi-signal conditional generation.

3481. PT-T2I/V: An Efficient Proxy-Tokenized Diffusion Transformer for Text-to-Image/Video-Task

链接：https://iclr.cc/virtual/2025/poster/28517 abstract： The global self-attention mechanism in diffusion transformers involves redundant computation due to the sparse and redundant nature of visual information, and the attention map of tokens within a spatial window shows significant similarity. To address this redundancy, we propose the Proxy-Tokenized Diffusion Transformer (PT-DiT), which employs sparse representative token attention (where the number of representative tokens is much smaller than the total number of tokens) to efficiently model global visual information. Specifically, within each transformer block, we compute an averaging token from each spatial-temporal window to serve as a proxy token for that region. The global semantics are captured through the self-attention of these proxy tokens and then injected into all latent tokens via cross-attention. Simultaneously, we introduce window and shift window attention to address the limitations in detail modeling caused by the sparse attention mechanism. Building on the well-designed PT-DiT, we further develop the PT-T2I/V family, which includes a variety of models for T2I, T2V, and T2MV tasks. Experimental results show that PT-DiT achieves competitive performance while reducing computational complexity in image and video generation tasks (e.g., a reduction 59\% compared to DiT and a reduction 34\% compared to PixArt-α). The visual exhibition of and code are available at https://360cvgroup.github.io/Qihoo-T2X/.

3482. ODE-based Smoothing Neural Network for Reinforcement Learning Tasks

链接：https://iclr.cc/virtual/2025/poster/29629 abstract： The smoothness of control actions is a significant challenge faced by deep reinforcement learning (RL) techniques in solving optimal control problems. Existing RL-trained policies tend to produce non-smooth actions due to high-frequency input noise and unconstrained Lipschitz constants in neural networks. This article presents a Smooth ODE (SmODE) network capable of simultaneously addressing both causes of unsmooth control actions, thereby enhancing policy performance and robustness under noise condition. We first design a smooth ODE neuron with first-order low-pass filtering expression, which can dynamically filter out high frequency noises of hidden state by a learnable state-based system time constant. Additionally, we construct a state-based mapping function, g, and theoretically demonstrate its capacity to control the ODE neuron's Lipschitz constant. Then, based on the above neuronal structure design, we further advanced the SmODE network serving as RL policy approximators. This network is compatible with most existing RL algorithms, offering improved adaptability compared to prior approaches. Various experiments show that our SmODE network demonstrates superior anti-interference capabilities and smoother action outputs than the multi-layer perception and smooth network architectures like LipsNet.

3483. Efficient Learning with Sine-Activated Low-Rank Matrices

链接：https://iclr.cc/virtual/2025/poster/29043 abstract： Low-rank decomposition has emerged as a vital tool for enhancing parameter efficiency in neural network architectures, gaining traction across diverse applications in machine learning. These techniques significantly lower the number of parameters, striking a balance between compactness and performance. However, a common challenge has been the compromise between parameter efficiency and the accuracy of the model, where reduced parameters often lead to diminished accuracy compared to their full-rank counterparts. In this work, we propose a novel theoretical framework that integrates a sinusoidal function within the low-rank decomposition process. This approach not only preserves the benefits of the parameter efficiency characteristic of low-rank methods but also increases the decomposition's rank, thereby enhancing model performance. Our method proves to be a plug in enhancement for existing low-rank models, as evidenced by its successful application in Vision Transformers (ViT), Large Language Models (LLMs), Neural Radiance Fields (NeRF) and 3D shape modelling.

3484. Model Editing as a Robust and Denoised variant of DPO: A Case Study on Toxicity

链接：https://iclr.cc/virtual/2025/poster/28521 abstract： Recent alignment algorithms such as direct preference optimization (DPO) have been developed to improve the safety of large language models (LLMs) by training these models to match human behaviors exemplified by preference data. However, these methods are both computationally intensive and lacking in controllability and transparency, inhibiting their widespread use. Furthermore, these tuning-based methods require large-scale preference data for training and are susceptible to noisy preference data. In this paper, we introduce a tuning-free alignment alternative, ProFS (Projection Filter for Subspaces), and demonstrate its effectiveness under the use case of toxicity reduction. Grounded on theory from factor analysis, ProFS is a sample-efficient model editing approach that identifies a toxic subspace in the model parameter space and reduces model toxicity by projecting away the detected subspace. The toxic subspace is identified by extracting preference data embeddings from the language model, and removing non-toxic information from these embeddings. We show that ProFS is more sample-efficient than DPO, further showcasing greater robustness to noisy data. Finally, we attempt to connect tuning based alignment with editing, by establishing both theoretical and empirical connections between ProFS and DPO, showing that ProFS can be interpreted as a denoised version of a single DPO step.

3485. Group Distributionally Robust Dataset Distillation with Risk

Minimization

链接：https://iclr.cc/virtual/2025/poster/31083 abstract： Dataset distillation (DD) has emerged as a widely adopted technique for crafting a synthetic dataset that captures the essential information of a training dataset, facilitating the training of accurate neural models. Its applications span various domains, including transfer learning, federated learning, and neural architecture search. The most popular methods for constructing the synthetic data rely on matching the convergence properties of training the model with the synthetic dataset and the training dataset. However, using the empirical loss as the criterion must be thought of as auxiliary in the same sense that the training set is an approximate substitute for the population distribution, and the latter is the data of interest. Yet despite its popularity, an aspect that remains unexplored is the relationship of DD to its generalization, particularly across uncommon subgroups. That is, how can we ensure that a model trained on the synthetic dataset performs well when faced with samples from regions with low population density? Here, the representativeness and coverage of the dataset become salient over the guaranteed training error at inference. Drawing inspiration from distributionally robust optimization, we introduce an algorithm that combines clustering with the minimization of a risk measure on the loss to conduct DD. We provide a theoretical rationale for our approach and demonstrate its effective generalization and robustness across subgroups through numerical experiments.

3486. Proactive Privacy Amnesia for Large Language Models: Safeguarding PII with Negligible Impact on Model Utility

链接：https://iclr.cc/virtual/2025/poster/28672 abstract： With the rise of large language models (LLMs), increasing research has recognizedtheir risk of leaking personally identifiable information (PII) under maliciousattacks. Although efforts have been made to protect PII in LLMs, existing methodsstruggle to balance privacy protection with maintaining model utility. In this paper,inspired by studies of amnesia in cognitive science, we propose a novel approach,Proactive Privacy Amnesia (PPA), to safeguard PII in LLMs while preserving theirutility. This mechanism works by actively identifying and forgetting key memoriesmost closely associated with PII in sequences, followed by a memory implantingusing suitable substitute memories to maintain the LLM’s functionality. We conductevaluations across multiple models to protect common PII, such as phone numbersand physical addresses, against prevalent PII-targeted attacks, demonstrating thesuperiority of our method compared with other existing defensive techniques. Theresults show that our PPA method completely eliminates the risk of phone numberexposure by 100% and significantly reduces the risk of physical address exposureby 9.8% – 87.6%, all while maintaining comparable model utility performance.

3487. Can In-context Learning Really Generalize to Out-of-distribution Tasks?

链接：https://iclr.cc/virtual/2025/poster/30175 abstract： In this work, we explore the mechanism of in-context learning (ICL) on out-of-distribution (OOD) tasks that were not encountered during training. To achieve this, we conduct synthetic experiments where the objective is to learn OOD mathematical functions through ICL using a GPT-2 model. We reveal that Transformers may struggle to learn OOD task functions through ICL. Specifically, ICL performance resembles implementing a function within the pretraining hypothesis space and optimizing it with gradient descent based on the in-context examples. Additionally, we investigate ICL's well-documented ability to learn unseen abstract labels in context. We demonstrate that such ability only manifests in the scenarios without distributional shifts and, therefore, may not serve as evidence of new-task-learning ability. Furthermore, we assess ICL's performance on OOD tasks when the model is pretrained on multiple tasks. Both empirical and theoretical analyses demonstrate the existence of the \textbf{low-test-error preference} of ICL, where it tends to implement the pretraining function that yields low test error in the testing context. We validate this through numerical experiments. This new theoretical result, combined with our empirical findings, elucidates the mechanism of ICL in addressing OOD tasks.

3488. Rethinking Invariance in In-context Learning

链接：https://iclr.cc/virtual/2025/poster/28271 abstract：

3489. What is Wrong with Perplexity for Long-context Language Modeling?

链接：https://iclr.cc/virtual/2025/poster/28884 abstract： Handling long-context inputs is crucial for large language models (LLMs) in tasks such as extended conversations, document summarization, and many-shot in-context learning. While recent approaches have extended the context windows of LLMs and employed perplexity (PPL) as a standard evaluation metric, PPL has proven unreliable for assessing long-context capabilities. The underlying cause of this limitation has remained unclear. In this work, we provide a comprehensive explanation for this issue. We find that PPL overlooks key tokens, which are essential for long-context understanding, by averaging across all tokens and thereby obscuring the true performance of models in long-context scenarios. To address this, we propose \textbf{LongPPL}, a novel metric that focuses on key tokens by employing a long-short context contrastive method to identify them. Our experiments demonstrate that LongPPL strongly correlates with performance on various long-context benchmarks (e.g., Pearson correlation of -0.96), significantly outperforming traditional PPL in predictive accuracy. Additionally, we introduce \textbf{LongCE} (Long-context Cross-Entropy) loss, a re-weighting strategy for fine-tuning that prioritizes key tokens, leading to consistent improvements across diverse benchmarks. In summary, these contributions offer deeper insights into the limitations of PPL and present effective solutions for accurately evaluating and enhancing the long-context capabilities of LLMs. Code is available at https://github.com/PKU-ML/LongPPL.

3490. Scaling Offline Model-Based RL via Jointly-Optimized World-Action Model Pretraining

链接：https://iclr.cc/virtual/2025/poster/29567 abstract： A significant aspiration of offline reinforcement learning (RL) is to develop a generalist agent with high capabilities from large and heterogeneous datasets. However, prior approaches that scale offline RL either rely heavily on expert trajectories or struggle to generalize to diverse unseen tasks. Inspired by the excellent generalization of world model in conditional video generation, we explore the potential of image observation-based world model for scaling offline RL and enhancing generalization on novel tasks. In this paper, we introduce JOWA: Jointly-Optimized World-Action model, an offline model-based RL agent pretrained on multiple Atari games with 6 billion tokens data to learn general-purpose representation and decision-making ability. Our method jointly optimizes a world-action model through a shared transformer backbone, which stabilize temporal difference learning with large models during pretraining. Moreover, we propose a provably efficient and parallelizable planning algorithm to compensate for the Q-value estimation error and thus search out better policies. Experimental results indicate that our largest agent, with 150 million parameters, achieves 78.9% human-level performance on pretrained games using only 10% subsampled offline data, outperforming existing state-of-the-art large-scale offline RL baselines by 31.6% on averange. Furthermore, JOWA scales favorably with model capacity and can sample-efficiently transfer to novel games using only 5k offline fine-tuning data (approximately 4 trajectories) per game, demonstrating superior generalization.

3491. Conformal Prediction Sets Can Cause Disparate Impact

链接：https://iclr.cc/virtual/2025/poster/28872 abstract： Conformal prediction is a statistically rigorous method for quantifying uncertainty in models by having them output sets of predictions, with larger sets indicating more uncertainty. However, prediction sets are not inherently actionable; many applications require a single output to act on, not several. To overcome this limitation, prediction sets can be provided to a human who then makes an informed decision. In any such system it is crucial to ensure the fairness of outcomes across protected groups, and researchers have proposed that Equalized Coverage be used as the standard for fairness. By conducting experiments with human participants, we demonstrate that providing prediction sets can lead to disparate impact in decisions. Disquietingly, we find that providing sets that satisfy Equalized Coverage actually increases disparate impact compared to marginal coverage. Instead of equalizing coverage, we propose to equalize set sizes across groups which empirically leads to lower disparate impact.

3492. TFG-Flow: Training-free Guidance in Multimodal Generative Flow

链接：https://iclr.cc/virtual/2025/poster/30288 abstract： Given an unconditional generative model and a predictor for a target property (e.g., a classifier), the goal of training-free guidance is to generate samples with desirable target properties without additional training. As a highly efficient technique for steering generative models toward flexible outcomes, training-free guidance has gained increasing attention in diffusion models. However, existing methods only handle data in continuous spaces, while many scientific applications involve both continuous and discrete data (referred to as multimodality). Another emerging trend is the growing use of the simple and general flow matching framework in building generative foundation models, where guided generation remains under-explored. To address this, we introduce TFG-Flow, a novel training-free guidance method for multimodal generative flow. TFG-Flow addresses the curse-of-dimensionality while maintaining the property of unbiased sampling in guiding discrete variables. We validate TFG-Flow on four molecular design tasks and show that TFG-Flow has great potential in drug design by generating molecules with desired properties.

3493. Infinite-Resolution Integral Noise Warping for Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29273 abstract： Adapting pretrained image-based diffusion models to generate temporally consistent videos has become an impactful generative modeling research direction. Training-free noise-space manipulation has proven to be an effective technique, where the challenge is to preserve the Gaussian white noise distribution while adding in temporal consistency. Recently, Chang et al. (2024) formulated this problem using an integral noise representation with distribution-preserving guarantees, and proposed an upsampling-based algorithm to compute it. However, while their mathematical formulation is advantageous, the algorithm incurs a high computational cost. Through analyzing the limiting-case behavior of their algorithm as the upsampling resolution goes to infinity, we develop an alternative algorithm that, by gathering increments of multiple Brownian bridges, achieves their infinite-resolution accuracy while simultaneously reducing the computational cost by orders of magnitude. We prove and experimentally validate our theoretical claims, and demonstrate our method's effectiveness in real-world applications. We further show that our method can readily extend to the 3-dimensional space.

3494. Is Factuality Enhancement a Free Lunch For LLMs? Better Factuality Can Lead to Worse Context-Faithfulness

链接：https://iclr.cc/virtual/2025/poster/29143 abstract： As the modern tools of choice for text understanding and generation, large language models (LLMs) are expected to accurately output answers by leveraging the input context.This requires LLMs to possess both context-faithfulness and factual accuracy.While extensive efforts aim to reduce hallucinations through factuality enhancement methods, they also pose risks of hindering context-faithfulness, as factuality enhancement can lead LLMs to become overly confident in their parametric knowledge, causing them to overlook the relevant input context.In this work, we argue that current factuality enhancement methods can significantly undermine the context-faithfulness of LLMs.We first revisit the current factuality enhancement methods and evaluate their effectiveness in enhancing factual accuracy.Next, we evaluate their performance on knowledge editing tasks to assess the potential impact on context-faithfulness.The experimental results reveal that while these methods may yield inconsistent improvements in factual accuracy, they also cause a more severe decline in context-faithfulness, with the largest decrease reaching a striking 69.7\%.To explain these declines, we analyze the hidden states and logit distributions for the tokens representing new knowledge and parametric knowledge respectively, highlighting the limitations of current approaches.Our finding highlights the complex trade-offs inherent in enhancing LLMs.Therefore, we recommend that more research on LLMs' factuality enhancement make efforts to reduce the sacrifice of context-faithfulness.

3495. MeshAnything: Artist-Created Mesh Generation with Autoregressive Transformers

链接：https://iclr.cc/virtual/2025/poster/30068 abstract： Recently, 3D assets created via reconstruction and generation have matched the quality of manually crafted assets, highlighting their potential for replacement. However, this potential is largely unrealized because these assets always need to be converted to meshes for 3D industry applications, and the meshes produced by current mesh extraction methods are significantly inferior to Artist-Created Meshes (AMs), i.e., meshes created by human artists. Specifically, current mesh extraction methods rely on dense faces and ignore geometric features, leading to inefficiencies, complicated post-processing, and lower representation quality.To address these issues, we introduce MeshAnything, a model that treats mesh extraction as a generation problem, producing AMs aligned with specified shapes.By converting 3D assets in any 3D representation into AMs, MeshAnything can be integrated with various 3D asset production methods, thereby enhancing their application across the 3D industry.The architecture of MeshAnything comprises a VQ-VAE and a shape-conditioned decoder-only transformer. We first learn a mesh vocabulary using the VQ-VAE, then train the shape-conditioned decoder-only transformer on this vocabulary for shape-conditioned autoregressive mesh generation. Our extensive experiments show that our method generates AMs with hundreds of times fewer faces, significantly improving storage, rendering, and simulation efficiencies, while achieving precision comparable to previous methods.

3496. RazorAttention: Efficient KV Cache Compression Through Retrieval Heads

链接：https://iclr.cc/virtual/2025/poster/28028 abstract： The memory and computational demands of Key-Value (KV) cache present significant challenges for deploying long-context language models. Previous approaches attempt to mitigate this issue by selectively dropping tokens, which irreversibly erases critical information that might be needed for future queries. In this paper, we propose a novel compression technique for KV cache that preserves all token information. Our investigation reveals that: i) Most attention heads primarily focus on the local context; ii) Only a few heads, denoted as retrieval heads, can essentially pay attention to all input tokens. These key observations motivate us to use separate caching strategy for attention heads.Therefore, we propose RazorAttention, a training-free KV cache compression algorithm, which maintains a full cache for these crucial retrieval heads and discards the remote tokens in non-retrieval heads. Furthermore, we introduce a novel mechanism involving a “compensation token” to further recover the information in the dropped tokens. Extensive evaluations across a diverse set of large language models (LLMs) demonstrate that RazorAttention achieves a reduction in KV cache size by over 70% without noticeable impacts on performance. Additionally, RazorAttention is compatible with FlashAttention, rendering it an efficient and plug-and-play solution that enhances LLM inference efficiency without overhead or retraining of the original model.

3497. Residual Kernel Policy Network: Enhancing Stability and Robustness in RKHS-Based Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/31103 abstract： Achieving optimal performance in reinforcement learning requires robust policies supported by training processes that ensure both sample efficiency and stability. Modeling the policy in reproducing kernel Hilbert space (RKHS) enables efficient exploration of local optimal solutions. However, the stability of existing RKHS-based methods is hindered by significant variance in gradients, while the robustness of the learned policies is often compromised due to the sensitivity of hyperparameters. In this work, we conduct a comprehensive analysis of the significant instability in RKHS policies and reveal that the variance of the policy gradient increases substantially when a wide-bandwidth kernel is employed. To address these challenges, we propose a novel RKHS policy learning method integrated with representation learning to dynamically process observations in complex environments, enhancing the robustness of RKHS policies. Furthermore, inspired by the advantage functions, we introduce a residual layer that further stabilizes the training process by significantly reducing gradient variance in RKHS. Our novel algorithm, the Residual Kernel Policy Network (ResKPN), demonstrates state-of-the-art performance, achieving a 30% improvement in episodic rewards across complex environments.

3498. Unposed Sparse Views Room Layout Reconstruction in the Age of Pretrain Model

链接：https://iclr.cc/virtual/2025/poster/30437 abstract：　Room layout estimation from multiple-perspective images is poorly investigated due to the complexities that emerge from multi-view geometry, which requires muti-step solutions such as camera intrinsic and extrinsic estimation, image matching, and triangulation. However, in 3D reconstruction, the advancement of recent 3D foundation models such as DUSt3R has shifted the paradigm from the traditional multi-step structure-from-motion process to an end-to-end single-step approach.To this end, we introduce Plane-DUSt3R, a novel method for multi-view room layout estimation leveraging the 3D foundation model DUSt3R. Plane-DUSt3R incorporates the DUSt3R framework and fine-tunes on a room layout dataset (Structure3D) with a modified objective to estimate structural planes. By generating uniform and parsimonious results, Plane-DUSt3R enables room layout estimation with only a single post-processing step and 2D detection results. Unlike previous methods that rely on single-perspective or panorama image, Plane-DUSt3R extends the setting to handle multiple-perspective images. Moreover, it offers a streamlined, end-to-end solution that simplifies the process and reduces error accumulation.Experimental results demonstrate that Plane-DUSt3R not only outperforms state-of-the-art methods on the synthetic dataset but also proves robust and effective on in the wild data with different image styles such as cartoon. Our code is available at: https://github.com/justacar/Plane-DUSt3R

3499. Unifying Unsupervised Graph-Level Anomaly Detection and Out-of-Distribution Detection: A Benchmark

链接：https://iclr.cc/virtual/2025/poster/28832 abstract：　To build safe and reliable graph machine learning systems, unsupervised graph-level anomaly detection (GLAD) and unsupervised graph-level out-of-distribution (OOD) detection (GLOD) have received significant attention in recent years. Though these two lines of research share the same objective, they have been studied independently in the community due to distinct evaluation setups, creating a gap that hinders the application and evaluation of methods from one to the other. To bridge the gap, in this work, we present a Unified Benchmark for unsupervised Graph-level OOD and anomaly Detection (UB-GOLD), a comprehensive evaluation framework that unifies GLAD and GLOD under the concept of generalized graph-level OOD detection. Our benchmark encompasses 35 datasets spanning four practical anomaly and OOD detection scenarios, facilitating the comparison of 18 representative GLAD/GLOD methods. We conduct multi-dimensional analyses to explore the effectiveness, generalizability, robustness, and efficiency of existing methods, shedding light on their strengths and limitations. Furthermore, we provide an open-source codebase of UB-GOLD to foster reproducible research and outline potential directions for future investigations based on our insights.

3500. Doubly robust identification of treatment effects from multiple environments

链接：https://iclr.cc/virtual/2025/poster/30659 abstract：

3501. Can We Trust Embodied Agents? Exploring Backdoor Attacks against Embodied LLM-Based Decision-Making Systems

链接：https://iclr.cc/virtual/2025/poster/29631 abstract：　Large Language Models (LLMs) have shown significant promise in real-world decision-making tasks for embodied artificial intelligence, especially when fine-tuned to leverage their inherent common sense and reasoning abilities while being tailored to specific applications. However, this fine-tuning process introduces considerable safety and security vulnerabilities, especially in safety-critical cyber-physical systems. In this work, we propose the first comprehensive framework for Backdoor Attacks against LLM-based Decision-making systems (BALD) in embodied AI, systematically exploring the attack surfaces and trigger mechanisms. Specifically, we propose three distinct attack mechanisms: word injection, scenario manipulation, and knowledge injection, targeting various components in the LLM-based decision-making pipeline. We perform extensive experiments on representative LLMs (GPT-3.5, LLaMA2, PaLM2) in autonomous driving and home robot tasks, demonstrating the effectiveness and stealthiness of our backdoor triggers across various attack channels, with cases like vehicles accelerating toward obstacles and robots placing knives on beds. Our word and knowledge injection attacks achieve nearly 100\% success rate across multiple models and datasets while requiring only limited access to the system. Our scenario manipulation attack yields success rates exceeding 65\%, reaching up to 90\%, and does not require any runtime system intrusion. We also assess the robustness of these attacks against defenses, revealing their resilience. Our findings highlight critical security vulnerabilities in embodied LLM systems and emphasize the urgent need for safeguarding these systems to mitigate potential risks.

3502. Enhancing Federated Domain Adaptation with Multi-Domain Prototype-Based Federated Fine-Tuning

链接：https://iclr.cc/virtual/2025/poster/31049 abstract：　Federated Domain Adaptation (FDA) is a Federated Learning (FL) scenario where models are trained across multiple clients with unique data domains but a shared category space, without transmitting private data. The primary challenge in FDA is data heterogeneity, which causes significant divergences in gradient updates when using conventional averaging-based aggregation methods, reducing the efficacy of the global model. This further undermines both in-domain and out-of-domain performance (within the same federated system but outside the local client), which is critical in certain business applications. To address this, we propose a novel framework called \textbf{M}ulti-domain \textbf{P}rototype-based \textbf{F}ederated Fine-\textbf{T}uning (MPFT). MPFT fine-tunes a pre-trained model using multi-domain prototypes, i.e., several pretrained representations enriched with domain-specific information from category-specific local data. This enables supervised learning on the server to create a globally optimized adapter that is subsequently distributed to local clients, without the intrusion of data privacy. Empirical results show that MPFT significantly improves both in-domain and out-of-domain accuracy over conventional methods, enhancing knowledge preservation and adaptation in FDA. Notably, MPFT achieves convergence within a single communication round, greatly reducing computation and communication costs. To ensure privacy, MPFT applies differential privacy to protect the prototypes. Additionally, we develop a prototype-based feature space hijacking attack to evaluate robustness, confirming that raw data samples remain unrecoverable even after extensive training epochs. The complete implementation of MPFL is available at \url{https://anonymous.4open.science/r/DomainFL/}.

3503. Re-Imagining Multimodal Instruction Tuning: A Representation View

链接：https://iclr.cc/virtual/2025/poster/27640 abstract： Multimodal instruction tuning has proven to be an effective strategy for achieving zero-shot generalization by fine-tuning pre-trained Large Multimodal Models (LMMs) with instruction-following data. However, as the scale of LMMs continues to grow, fully fine-tuning these models has become highly parameter-intensive. Although Parameter-Efficient Fine-Tuning (PEFT) methods have been introduced to reduce the number of tunable parameters, a significant performance gap remains compared to full fine-tuning. Furthermore, existing PEFT approaches are often highly parameterized, making them difficult to interpret and control. In light of this, we introduce Multimodal Representation Tuning (MRT), a novel approach that focuses on directly editing semantically rich multimodal representations to achieve strong performance and provide intuitive control over LMMs. Empirical results show that our method surpasses current state-of-the-art baselines with significant performance gains (e.g., 1580.40 MME score) while requiring substantially fewer tunable parameters (e.g., 0.03% parameters). Additionally, we conduct experiments on editing instrumental tokens within multimodal representations, demonstrating that direct manipulation of these representations enables simple yet effective control over network behavior.

3504. Graph-Guided Scene Reconstruction from Images with 3D Gaussian Splatting

链接：https://iclr.cc/virtual/2025/poster/30969 abstract： This paper investigates an open research challenge of reconstructing high-quality, large-scale 3D open scenes from images. It is observed existing methods have various limitations, such as requiring precise camera poses for input and dense viewpoints for supervision. To perform effective and efficient 3D scene reconstruction, we propose a novel graph-guided 3D scene reconstruction framework, GraphGS. Specifically, given a set of images captured by RGB cameras on a scene, we first design a spatial prior-based scene structure estimation method. This is then used to create a camera graph that includes information about the camera topology. Further, we propose to apply the graph-guided multi-view consistency constraint and adaptive sampling strategy to the 3D Gaussian Splatting optimization process. This greatly alleviates the issue of Gaussian points overfitting to specific sparse viewpoints and expedites the 3D reconstruction process. We demonstrate GraphGS achieves high-fidelity 3D reconstruction from images, which presents state-of-the-art performance through quantitative and qualitative evaluation across multiple datasets.

3505. Dynamic Diffusion Transformer

链接：https://iclr.cc/virtual/2025/poster/28043 abstract： Diffusion Transformer (DiT), an emerging diffusion model for image generation,has demonstrated superior performance but suffers from substantial computationalcosts. Our investigations reveal that these costs stem from the static inferenceparadigm, which inevitably introduces redundant computation in certain diffusiontimesteps and spatial regions. To address this inefficiency, we propose DynamicDiffusion Transformer (DyDiT), an architecture that dynamically adjusts its compu-tation along both timestep and spatial dimensions during generation. Specifically,we introduce a Timestep-wise Dynamic Width (TDW) approach that adapts modelwidth conditioned on the generation timesteps. In addition, we design a Spatial-wise Dynamic Token (SDT) strategy to avoid redundant computation at unnecessaryspatial locations. Extensive experiments on various datasets and different-sizedmodels verify the superiority of DyDiT. Notably, with <3% additional fine-tuning it-erations, our method reduces the FLOPs of DiT-XL by 51%, accelerates generationby 1.73×, and achieves a competitive FID score of 2.07 on ImageNet.

3506. Bridging the Gap between Database Search and \emph{De Novo} Peptide Sequencing with SearchNovo

链接：https://iclr.cc/virtual/2025/poster/29582 abstract： Accurate protein identification from mass spectrometry (MS) data is fundamental to unraveling the complex roles of proteins in biological systems, with peptide sequencing being a pivotal step in this process. The two main paradigms for peptide sequencing are database search, which matches experimental spectra with peptide sequences from databases, and \emph{de novo} sequencing, which infers peptide sequences directly from MS without relying on pre-constructed database. Although database search methods are highly accurate, they are limited by their inability to identify novel, modified, or mutated peptides absent from the database. In contrast, \emph{de novo} sequencing is adept at discovering novel peptides but often struggles with missing peaks issue, further leading to lower precision. We introduce SearchNovo, a novel framework that synergistically integrates the strengths of database search and \emph{de novo} sequencing to enhance peptide sequencing. SearchNovo employs an efficient search mechanism to retrieve the most similar peptide spectrum match (PSM) from a database for each query spectrum, followed by a fusion module that utilizes the reference peptide sequence to guide the generation of the target sequence. Furthermore, we observed that dissimilar (noisy) reference peptides negatively affect model performance. To mitigate this, we constructed pseudo reference PSMs to minimize their impact.

Comprehensive evaluations on multiple datasets reveal that SearchNovo significantly outperforms state-of-the-art models. Also, analysis indicates that many retrieved spectra contain missing peaks absent in the query spectra, and the retrieved reference peptides often share common fragments with the target peptides. These are key elements in the recipe for SearchNovo's success. The code is available at: \textcolor{magenta}{\url{https://github.com/junxia97/SearchNovo}}.

3507. ReNovo: Retrieval-Based \emph{De Novo} Mass Spectrometry Peptide Sequencing

链接：https://iclr.cc/virtual/2025/poster/27978 abstract： Proteomics is the large-scale study of proteins. Tandem mass spectrometry, as the only high-throughput technique for protein sequence identification, plays a pivotal role in proteomics research. One of the long-standing challenges in this field is peptide identification, which entails determining the specific peptide (sequence of amino acids) that corresponds to each observed mass spectrum. The conventional approach involves database searching, wherein the observed mass spectrum is scored against a pre-constructed peptide database. However, the reliance on pre-existing databases limits applicability in scenarios where the peptide is absent from existing databases. Such circumstances necessitate \emph{de novo} peptide sequencing, which derives peptide sequence solely from input mass spectrum, independent of any peptide database. Despite ongoing advancements in \emph{de novo} peptide sequencing, its performance still has considerable room for improvement, which limits its application in large-scale experiments. In this study, we introduce a novel \textbf{Re}trieval-based \emph{De \textbf{Novo}} peptide sequencing methodology, termed \textbf{ReNovo}, which draws inspiration from database search methods. Specifically, by constructing a datastore from training data, ReNovo can retrieve information from the datastore during the inference stage to conduct retrieval-based inference, thereby achieving improved performance. This innovative approach enables ReNovo to effectively combine the strengths of both methods: utilizing the assistance of the datastore while also being capable of predicting novel peptides that are not present in pre-existing databases. A series of experiments have confirmed that ReNovo outperforms state-of-the-art models across multiple widely-used datasets, incurring only minor storage and time consumption, representing a significant advancement in proteomics. Supplementary materials include the code.

3508. Lotus: Diffusion-based Visual Foundation Model for High-quality Dense Prediction

链接：https://iclr.cc/virtual/2025/poster/28092 abstract： Leveraging the visual priors of pre-trained text-to-image diffusion models offers a promising solution to enhance zero-shot generalization in dense prediction tasks. However, existing methods often uncritically use the original diffusion formulation, which may not be optimal due to the fundamental differences between dense prediction and image generation. In this paper, we provide a systemic analysis of the diffusion formulation for the dense prediction, focusing on both quality and efficiency. And we find that the original parameterization type for image generation, which learns to predict noise, is harmful for dense prediction; the multi-step noising/denoising diffusion process is also unnecessary and challenging to optimize. Based on these insights, we introduce $\textbf{Lotus}$, a diffusion-based visual foundation model with a simple yet effective adaptation protocol for dense prediction. Specifically, Lotus is trained to directly predict annotations instead of noise, thereby avoiding harmful variance. We also reformulate the diffusion process into a single-step procedure, simplifying optimization and significantly boosting inference speed. Additionally, we introduce a novel tuning strategy called detail preserver, which achieves more accurate and fine-grained predictions. Without scaling up the training data or model capacity, Lotus achieves SoTA performance in zero-shot depth and normal estimation across various datasets. It also enhances efficiency, being significantly faster than most existing diffusion-based methods. Lotus' superior quality and efficiency also enable a wide range of practical applications, such as joint estimation, single/multi-view 3D reconstruction, etc.

3509. Factor Graph-based Interpretable Neural Networks

链接：https://iclr.cc/virtual/2025/poster/31227 abstract： Comprehensible neural network explanations are foundations for a better understanding of decisions, especially when the input data are infused with malicious perturbations. Existing solutions generally mitigate the impact of perturbations through adversarial training, yet they fail to generate comprehensible explanations under unknown perturbations. To address this challenge, we propose AGAIN, a factor graph-based interpretable neural network, which is capable of generating comprehensible explanations under unknown perturbations. Instead of retraining like previous solutions, the proposed AGAIN directly integrates logical rules by which logical errors in explanations are identified and rectified during inference. Specifically, we construct the factor graph to express logical rules between explanations and categories. By treating logical rules as exogenous knowledge, AGAIN can identify incomprehensible explanations that violate real-world logic. Furthermore, we propose an interactive intervention switch strategy rectifying explanations based on the logical guidance from the factor graph without learning perturbations, which overcomes the inherent limitation of adversarial training-based methods in defending only against known perturbations. Additionally, we theoretically demonstrate the effectiveness of employing factor graph by proving that the comprehensibility of explanations is strongly correlated with factor graph. Extensive experiments are conducted on three datasets and experimental results illustrate the superior performance of AGAIN compared to state-of-the-art baselines.

3510. The adaptive complexity of parallelized log-concave sampling

链接：https://iclr.cc/virtual/2025/poster/30393 abstract： In large-data applications, such as the inference process of diffusion models, it is desirable to design sampling algorithms with a high degree of parallelization. In this work, we study the adaptive complexity of sampling, which is the minimum number of sequential rounds required to achieve sampling given polynomially many queries executed in parallel at each round. For unconstrained sampling, we examine distributions that are log-smooth or log-Lipschitz and log strongly or non-strongly concave. We show that an almost linear iteration algorithm cannot return a sample with a specific exponentially small error under total variation distance. For box-constrained sampling, we show that an almost linear iteration algorithm cannot return a sample with sup-polynomially small error under total variation distance for log-concave distributions. Our proof relies upon novel analysis with the characterization of the output for the hardness potentials based on the chain-like structure with random partition and classical smoothing techniques.

3511. Improved Approximation Algorithms for k-Submodular Maximization via Multilinear Extension

链接：https://iclr.cc/virtual/2025/poster/30407 abstract：

3512. Progressive Parameter Efficient Transfer Learning for Semantic Segmentation

链接：https://iclr.cc/virtual/2025/poster/29265 abstract： Parameter Efficient Transfer Learning (PETL) excels in downstream classification fine-tuning with minimal computational overhead, demonstrating its potential within the pre-train and fine-tune paradigm. However, recent PETL methods consistently struggle when fine-tuning for semantic segmentation tasks, limiting their broader applicability. In this paper, we identify that fine-tuning for semantic segmentation requires larger parameter adjustments due to shifts in semantic perception granularity. Current PETL approaches are unable to effectively accommodate these shifts, leading to significant performance degradation. To address this, we introduce ProPETL, a novel approach that incorporates an additional midstream adaptation to progressively align pre-trained models for segmentation tasks. Through this process, ProPETL achieves state-of-the-art performance on most segmentation benchmarks and, for the first time, surpasses full fine-tuning on the challenging COCO-Stuff10k dataset. Furthermore, ProPETL demonstrates strong generalization across various pre-trained models and scenarios, highlighting its effectiveness and versatility for broader adoption in segmentation tasks. Code is available at: https://github.com/weeknan/ProPETL.

3513. The Crucial Role of Samplers in Online Direct Preference Optimization

链接：https://iclr.cc/virtual/2025/poster/30363 abstract： Direct Preference Optimization (DPO) has emerged as a stable, scalable, and efficient solution for language model alignment.Despite its empirical success, the optimization properties, particularly the impact of samplers on its convergence rates, remain under-explored. In this paper, we provide a rigorous analysis of DPO's convergence rates with different sampling strategies under the exact gradient setting, revealing a surprising separation: uniform sampling achieves $\textbf{linear}$ convergence, while our proposed online sampler achieves $\textbf{quadratic}$ convergence. We further adapt the sampler to practical settings by incorporating posterior distributions and logit mixing, demonstrating improvements over previous methods. For example, it outperforms vanilla DPO by over 7.4% on Safe-RLHF dataset. Our results not only offer insights into the theoretical understanding of DPO but also pave the way for further algorithm designs.

3514. I2VControl-Camera: Precise Video Camera Control with Adjustable Motion Strength

链接：https://iclr.cc/virtual/2025/poster/30626 abstract： Video generation technologies are developing rapidly and have broad potential applications. Among these technologies, camera control is crucial for generating professional-quality videos that accurately meet user expectations. However, existing camera control methods still suffer from several limitations, including control precision and the neglect of the control for subject motion dynamics. In this work, we propose I2VControl-Camera, a novel camera control method that significantly enhances controllability while providing adjustability over the strength of subject motion. To improve control precision, we employ point trajectory in the camera coordinate system instead of only extrinsic matrix information as our control signal. To accurately control and adjust the strength of subject motion, we explicitly model the higher-order components of the video trajectory expansion, not merely the linear terms, and design an operator that effectively represents the motion strength. We use an adapter architecture that is independent of the base model structure. Experiments on static and dynamic scenes show that our framework outperformances previous methods both quantitatively and qualitatively. Project page: https://wanquanf.github.io/I2VControlCamera.

3515. Reinforcement Learning from Imperfect Corrective Actions and Proxy Rewards

链接：https://iclr.cc/virtual/2025/poster/30110 abstract： In practice, reinforcement learning (RL) agents are often trained with a possibly imperfect proxy reward function, which may lead to a human-agent alignment issue (i.e., the learned policy either converges to non-optimal performance with low cumulative rewards, or achieves high cumulative rewards but in an undesired manner). To tackle this issue, we consider a framework where a human labeler can provide additional feedback in the form of corrective actions, which expresses the labeler's action preferences although this feedback may possibly be imperfect as well. In this setting, to obtain a better-aligned policy guided by both learning signals, we propose a novel value-based deep RL algorithm called Iterative learning from Corrective actions and Proxy rewards (ICoPro), which cycles through three phases: (1) Solicit sparse corrective actions from a human labeler on the agent's demonstrated trajectories; (2) Incorporate these corrective actions into the Q-function using a margin loss to enforce adherence to labeler's preferences; (3) Train the agent with standard RL losses regularized with a margin loss to learn from proxy rewards and propagate the Q-values learned from human feedback. Moreover, another novel design in our approach is to integrate pseudo-labels from the target Q-network to reduce human labor and further stabilize training. We experimentally validate our proposition on a variety of tasks (Atari games and autonomous driving on highway). On the one hand, using proxy rewards with different levels of imperfection, our method can better align with human and is more sample-efficient than baseline methods. On the other hand, facing corrective actions with different types of imperfection, our method can overcome the non-optimality of this feedback thanks to the guidance from proxy rewards.

3516. The Crystal Ball Hypothesis in diffusion models: Anticipating object positions from initial noise

链接：https://iclr.cc/virtual/2025/poster/30257 abstract： Diffusion models have achieved remarkable success in text-to-image generation tasks, yet the influence of initial noise remains largely unexplored. In this study, we identify specific regions within the initial noise image, termed trigger patches, that play a key role in inducing object generation in the resulting images. Notably, these patches are universal and can be generalized across various positions, seeds, and prompts. To be specific, extracting these patches from one noise and injecting them into another noise leads to object generation in targeted areas. To identify the trigger patches even before the image has been generated, just like consulting the crystal ball to foresee fate, we first create a dataset consisting of Gaussian noises labeled with bounding boxes corresponding to the objects appearing in the generated images and train a detector that identifies these patches from the initial noise. To explain the formation of these patches, we reveal that they are outliers in Gaussian noise, and follow distinct distributions through two-sample tests. These outliers can take effect when injected into different noises and generalize well across different settings. Finally, we find the misalignment between prompts and the trigger patch patterns can result in unsuccessful image generations. To overcome it, we propose a reject-sampling strategy to obtain optimal noise, aiming to improve prompt adherence and positional diversity in image generation.

3517. BenTo: Benchmark Reduction with In-Context Transferability

链接：https://iclr.cc/virtual/2025/poster/28565 abstract： Evaluating large language models (LLMs) is costly: it requires the generation and examination of LLM outputs on a large-scale benchmark of various tasks. This paper investigates how to efficiently reduce the tasks used to benchmark LLMs without affecting the evaluation quality. Our study reveals that task transferability and relevance provide critical information to identify the most representative subset of tasks via optimizing a facility location function. We propose a practically efficient metric for estimating the transferability between two tasks via in-context learning (ICL). By analyzing the pairwise transferability, we can reduce tasks in a modern LLM benchmark (e.g., MMLU or FLAN) to 5\% while inducing only a <4\% difference to the evaluation on the original benchmark. Compared to prior works, our method is training-free, gradient-free, and highly efficient requiring ICL only.

3518. Is Your Multimodal Language Model Oversensitive to Safe Queries?

链接：https://iclr.cc/virtual/2025/poster/29670 abstract： Humans are prone to cognitive distortions — biased thinking patterns that lead to exaggerated responses to specific stimuli, albeit in very different contexts.This paper demonstrates that advanced Multimodal Large Language Models (MLLMs) exhibit similar tendencies.While these models are designed to respond queries under safety mechanism, they sometimes reject harmless queries in the presence of certain visual stimuli, disregarding the benign nature of their contexts.As the initial step in investigating this behavior, we identify three representative types of stimuli that trigger the oversensitivity of existing MLLMs: $\textbf{\textit{Exaggerated Risk}}$, $\textbf{\textit{Negated Harm}}$, and $\textbf{\textit{Counterintuitive Interpretation}}$.To systematically evaluate MLLMs' oversensitivity to these stimuli, we propose the $\textbf{M}$ultimodal $\textbf{O}$ver$\textbf{S}$en$\textbf{S}$itivity $\textbf{Bench}$mark (MOSSBench).This toolkit consists of 300 manually collected benign multimodal queries, cross-verified by third-party reviewers (AMT).Empirical studies using MOSSBench on 20 MLLMs reveal several insights:(1). Oversensitivity is prevalent among SOTA MLLMs, with refusal rates reaching up to $\textbf{76}$\% for harmless queries.(2). Safer models are more oversensitive: increasing safety may inadvertently raise caution and conservatism in the model’s responses.(3). Different types of stimuli tend to cause errors at specific stages — perception, intent reasoning, and safety judgement — in the response process of MLLMs.These findings highlight the need for refined safety mechanisms that balance caution with contextually appropriate responses, improving the reliability of MLLMs in real-world applications.

3519. Many-Objective Multi-Solution Transport

链接：https://iclr.cc/virtual/2025/poster/29867 abstract： Optimizing the performance of many objectives (instantiated by tasks or clients) jointly with a few Pareto stationary solutions (models) is critical in machine learning. However, previous multi-objective optimization methods often focus on a few objectives and cannot scale to many objectives that outnumber the solutions, leading to either subpar performance or ignored objectives. We introduce "Many-objective multi-solution Transport (MosT)", a framework that finds multiple diverse solutions in the Pareto front of many objectives. Our insight is to seek multiple solutions, each performing as a domain expert and focusing on a specific subset of objectives while collectively covering all of them. MosT formulates the problem as a bi-level optimization of weighted objectives for each solution, where the weights are defined by an optimal transport between objectives and solutions. Our algorithm ensures convergence to Pareto stationary solutions for complementary subsets of objectives. On a range of applications in federated learning, multi-task learning, and mixture-of-prompt learning for LLMs, MosT distinctly outperforms strong baselines, delivering high-quality, diverse solutions that profile the entire Pareto frontier, thus ensuring balanced trade-offs across many objectives.

3520. Earlier Tokens Contribute More: Learning Direct Preference Optimization From Temporal Decay Perspective

链接：https://iclr.cc/virtual/2025/poster/29798 abstract： Direct Preference Optimization (DPO) has gained attention as an efficient alternative to reinforcement learning from human feedback (RLHF) for aligning large language models (LLMs) with human preferences. Despite its advantages, DPO suffers from a length bias, generating responses longer than those from the reference model. Existing solutions like SimPO and SamPO address this issue but uniformly treat the contribution of rewards across sequences, overlooking temporal dynamics. To this end, we propose an enhanced preference optimization method that incorporates a temporal decay factor controlled by a gamma parameter. This dynamic weighting mechanism adjusts the influence of each reward based on its position in the sequence, prioritizing earlier tokens that are more critical for alignment. By adaptively focusing on more relevant feedback, our approach mitigates overfitting to less pertinent data and remains responsive to evolving human preferences. Experimental results on several benchmarks show that our approach consistently outperforms vanilla DPO by 5.9-8.8 points on AlpacaEval 2 and 3.3-9.7 points on Arena-Hard across different model architectures and sizes. Furthermore, additional experiments on mathematical and reasoning benchmarks (MMLU, GSM8K, and MATH) confirm that our method enhances performance without compromising general capabilities. Our codebase would be available at \url{https://github.com/LotuSrc/D2PO}.

3521. Integrating Protein Dynamics into Structure-Based Drug Design via Full-Atom Stochastic Flows

链接：https://iclr.cc/virtual/2025/poster/30663 abstract：

3522. Group Ligands Docking to Protein Pockets

链接：https://iclr.cc/virtual/2025/poster/27682 abstract： Molecular docking is a key task in computational biology that has attracted increasing interest from the machine learning community. While existing methods have achieved success, they generally treat each protein-ligand pair in isolation. Inspired by the biochemical observation that ligands binding to the same target protein tend to adopt similar poses, we propose \textsc{GroupBind}, a novel molecular docking framework that simultaneously considers multiple ligands docking to a protein. This is achieved by introducing an interaction layer for the group of ligands and a triangle attention module for embedding protein-ligand and group-ligand pairs. By integrating our approach with diffusion based docking model, we set a new state-of-the-art performance on the PDBBind blind docking benchmark, demonstrating the effectiveness of our paradigm in enhancing molecular docking accuracy.

3523. ProteinBench: A Holistic Evaluation of Protein Foundation Models

链接：https://iclr.cc/virtual/2025/poster/30553 abstract： Recent years have witnessed a surge in the development of protein foundation models, significantly improving performance in protein prediction and generative tasks ranging from 3D structure prediction and protein design to conformational dynamics. However, the capabilities and limitations associated with these models remain poorly understood due to the absence of a unified evaluation framework. To fill this gap, we introduce ProteinBench, a holistic evaluation framework designed to enhance the transparency of protein foundation models. Our approach consists of three key components: (i) A taxonomic classification of tasks that broadly encompass the main challenges in the protein domain, based on the relationships between different protein modalities; (ii) A multi-metric evaluation approach that assesses performance across four key dimensions: quality, novelty, diversity, and robustness; and (iii) In-depth analyses from various user objectives, providing a holistic view of model performance. Our comprehensive evaluation of protein foundation models reveals several key findings that shed light on their current capabilities and limitations. To promote transparency and facilitate further research, we release the evaluation dataset, code, and a public leaderboard publicly for further analysis and a general modular toolkit. We intend for ProteinBench to be a living benchmark for establishing a standardized, in-depth evaluation framework for protein foundation models, driving their development and application while fostering collaboration within the field.

3524. Enhancing Prediction Performance through Influence Measure

链接：https://iclr.cc/virtual/2025/poster/30037 abstract： In the field of machine learning, the pursuit of accurate models is ongoing. A key aspect of improving prediction performance lies in identifying which data points in the training set should be excluded and which high-quality, potentially unlabeled data points outside the training set should be incorporated to improve the model's performance on unseen data. To accomplish this, an effective metric is needed to evaluate the contribution of each data point toward enhancing overall model performance. This paper proposes the use of an influence measure as a metric to assess the impact of training data on test set performance. Additionally, we introduce a data selection method to optimize the training set as well as a dynamic active learning algorithm driven by the influence measure. The effectiveness of these methods is demonstrated through extensive simulations and real-world datasets.

3525. Point Cluster: A Compact Message Unit for Communication-Efficient Collaborative Perception

链接：https://iclr.cc/virtual/2025/poster/30972 abstract： The objective of the collaborative perception task is to enhance the individual agent's perception capability through message communication among neighboring agents. A central challenge lies in optimizing the inherent trade-off between perception ability and communication cost. To tackle this bottleneck issue, we argue that a good message unit should encapsulate both semantic and structural information in a sparse format, a feature not present in prior approaches. In this paper, we innovatively propose a compact message unit, namely point cluster, whose core idea is to represent potential objects efficiently with explicitly decoupled low-level structure information and high-level semantic information. Building upon this new message unit, we propose a comprehensive framework CPPC for communication-efficient collaborative perception. The core principle of CPPC is twofold: first, through strategical point sampling, structure information can be well preserved with a few key points, which can significantly reduce communication cost; second, the sequence format of point clusters enables efficient message aggregation by set matching and merging, thereby eliminating unnecessary computation generated when aligning squared BEV maps, especially for long-range collaboration. To handle time latency and pose errors encountered in real-world scenarios, we also carefully design parameter-free solutions that can adapt to different noisy levels without finetuning. Experiments on two widely recognized collaborative perception benchmarks showcase the superior performance of our method compared to the previous state-of-the-art approaches.

3526. Dense Video Object Captioning from Disjoint Supervision

链接：https://iclr.cc/virtual/2025/poster/29140 abstract： We propose a new task and model for dense video object captioning -- detecting, tracking and captioning trajectories of objects in a video. This task unifies spatial and temporal localization in video, whilst also requiring fine-grained visual understanding that is best described by natural language. We propose a unified model, and demonstrate how our end-to-end approach is more accurate and temporally coherent than a multi-stage pipeline combining state-of-the-art detection, tracking, and captioning models. Moreover, we propose a training strategy based on a mixture of disjoint tasks, which allows us to leverage diverse, large-scale datasets which supervise different parts of our model. Although each pretraining task only provides weak supervision, they are complementary and, when combined, result in noteworthy zero-shot ability and serve as strong initialization for additional finetuning to further improve accuracy. We carefully design new metrics capturing all components of our task, and show how we can repurpose existing video grounding datasets (e.g. VidSTG and VLN) for our new task. We show that our model improves upon a number of strong baselines for this new task. Furthermore, we can apply our model to the task of spatial grounding, outperforming prior state-of-the-art on VidSTG and VLN, without explicitly training for it. Our code is available at https://github.com/google-research/scenic.

3527. A Robust Method to Discover Causal or Anticausal Relation

链接：https://iclr.cc/virtual/2025/poster/29717 abstract： Understanding whether the data generative process follows causal or anticausal relations is important for many applications. Existing causal discovery methods struggle with high-dimensional perceptual data such as images. Moreover, they require well-labeled data, which may not be feasible due to measurement error. In this paper, we propose a robust method to detect whether the data generative process is causal or anticausal. To determine the causal or anticausal relation, we identify an asymmetric property: under the causal relation, the instance distribution does not contain information about the noisy class-posterior distribution. We also propose a practical method to verify this via a noise injection approach. Our method is robust to label errors and is designed to handle both large-scale and high-dimensional datasets effectively. Both theoretical analyses and empirical results on a variety of datasets demonstrate the effectiveness of our proposed method in determining the causal or anticausal direction of the data generative process.

3528. N-ForGOT: Towards Not-forgetting and Generalization of Open Temporal Graph Learning

链接：https://iclr.cc/virtual/2025/poster/28202 abstract： Temporal Graph Neural Networks (TGNNs) lay emphasis on capturing node interactions over time but often overlook evolution in node classes and dynamic data distributions triggered by the continuous emergence of new class labels, known as the open-set problem. This problem poses challenges for existing TGNNs in preserving learned classes while rapidly adapting to new, unseen classes. To address this, this paper identifies two primary factors affecting model performance on the open temporal graph, backed by a theoretical guarantee: (1) the forgetting of prior knowledge and (2) distribution discrepancies between successive tasks. Building on these insights, we propose N-ForGOT, which incorporates two plug-in modules into TGNNs to preserve prior knowledge and enhance model generalizability for new classes simultaneously. The first module preserves previously established inter-class connectivity and decision boundaries during the training of new classes to mitigate the forgetting caused by temporal evolutions of class characteristics. The second module introduces an efficient method for measuring distribution discrepancies with designed temporal Weisfeiler-Lehman subtree patterns, effectively addressing both structural and temporal shifts while reducing time complexity. Experimental results on four public datasets demonstrate that our method significantly outperforms state-of-the-art approaches in prediction accuracy, prevention of forgetting, and generalizability.

3529. HiSplat: Hierarchical 3D Gaussian Splatting for Generalizable Sparse-View Reconstruction

链接：https://iclr.cc/virtual/2025/poster/29622 abstract： Reconstructing 3D scenes from multiple viewpoints is a fundamental task in stereo vision. Recently, advances in generalizable 3D Gaussian Splatting have enabled high-quality novel view synthesis for unseen scenes from sparse input views by feed-forward predicting per-pixel Gaussian parameters without extra optimization. However, existing methods typically generate single-scale 3D Gaussians, which lack representation of both large-scale structure and texture details, resulting in mislocation and artefacts. In this paper, we propose a novel framework, HiSplat, which introduces a hierarchical manner in generalizable 3D Gaussian Splatting to construct hierarchical 3D Gaussians via a coarse-to-fine strategy. Specifically, HiSplat generates large coarse-grained Gaussians to capture large-scale structures, followed by fine-grained Gaussians to enhance delicate texture details. To promote inter-scale interactions, we propose an Error Aware Module for Gaussian compensation and a Modulating Fusion Module for Gaussian repair. Our method achieves joint optimization of hierarchical representations, allowing for novel view synthesis using only two-view reference images. Comprehensive experiments on various datasets demonstrate that HiSplat significantly enhances reconstruction quality and cross-dataset generalization compared to prior single-scale methods. The corresponding ablation study and analysis of different-scale 3D Gaussians reveal the mechanism behind the effectiveness. Code is at https://github.com/Open3DVLab/HiSplat.

3530. UniGS: Unified Language-Image-3D Pretraining with Gaussian Splatting

链接：https://iclr.cc/virtual/2025/poster/30882 abstract： Recent advancements in multi-modal 3D pre-training methods have shown promising efficacy in learning joint representations of text, images, and point clouds. However, adopting point clouds as 3D representation fails to fully capture the intricacies of the 3D world and exhibits a noticeable gap between the discrete points and the dense 2D pixels of images. To tackle this issue, we propose UniGS, integrating 3D Gaussian Splatting (3DGS) into multi-modal pre-training to enhance the 3D representation. We first rely on the 3DGS representation to model the 3D world as a collection of 3D Gaussians with color and opacity, incorporating all the information of the 3D scene while establishing a strong connection with 2D images. Then, to achieve Language-Image-3D pertaining, UniGS starts with a pretrained vision-language model to establish a shared visual and textual space through extensive real-world image-text pairs. Subsequently, UniGS employs a 3D encoder to align the optimized 3DGS with the Language-Image representations to learn unified multi-modal representations. To facilitate the extraction of global explicit 3D features by the 3D encoder and achieve better cross-modal alignment, we additionally introduce a novel Gaussian-Aware Guidance module that guides the learning of fine-grained representations of the 3D domain. Through extensive experiments across the Objaverse, ABO, MVImgNet and SUN RGBD datasets with zero-shot classification, text-driven retrieval and open-world understanding tasks, we demonstrate the effectiveness of UniGS in learning a more general and stronger aligned multi-modal representation. Specifically, UniGS achieves leading results across different 3D tasks with remarkable improvements over previous SOTA, Uni3D, including on zero-shot classification (+9.36%), text-driven retrieval (+4.3%) and open-world understanding (+7.92%).

3531. CryoFM: A Flow-based Foundation Model for Cryo-EM Densities

链接：https://iclr.cc/virtual/2025/poster/29563 abstract： Cryo-electron microscopy (cryo-EM) is a powerful technique in structural biology and drug discovery, enabling the study of biomolecules at high resolution. Significant advancements by structural biologists using cryo-EM have led to the production of around 40k protein density maps at various resolutions. However, cryo-EM data processing algorithms have yet to fully benefit from our knowledge of biomolecular density maps, with only a few recent models being data-driven but limited to specific tasks. In this study, we present CryoFM, a foundation model designed as a generative model, learning the distribution of high-quality density maps and generalizing effectively to downstream tasks. Built on flow matching, CryoFM is trained to accurately capture the prior distribution of biomolecular density maps. Furthermore, we introduce a flow posterior sampling method that leverages CryoFM as a flexible prior for several downstream tasks in cryo-EM and cryo-electron tomography (cryo-ET) without the need for fine-tuning, achieving state-of-the-art performance on most tasks and demonstrating its potential as a foundational model for broader applications in these fields.

3532. Anyprefer: An Agentic Framework for Preference Data Synthesis

链接：https://iclr.cc/virtual/2025/poster/29342 abstract： High-quality preference data is essential for aligning foundation models with human values through preference learning. However, manual annotation of such data is often time-consuming and costly. Recent methods often adopt a self-rewarding approach, where the target model generates and annotates its own preference data, but this can lead to inaccuracies since the reward model shares weights with the target model, thereby amplifying inherent biases. To address these issues, we propose Anyprefer, a framework designed to synthesize high-quality preference data for aligning the target model. Anyprefer frames the data synthesis process as a cooperative two-player Markov Game, where the target model and the judge model collaborate together. Here, a series of external tools are introduced to assist the judge model in accurately rewarding the target model’s responses, mitigating biases in the rewarding process. In addition, a feedback mechanism is introduced to optimize prompts for both models, enhancing collaboration and improving data quality. The synthesized data is compiled into a new preference dataset, Anyprefer-V1, consisting of 58K high-quality preference pairs. Extensive experiments show that Anyprefer significantly improves model alignment performance across four main applications, covering 21 datasets, achieving average improvements of 18.55% in five natural language generation datasets, 3.66% in nine vision-language understanding datasets, 30.05% in three medical image analysis datasets, and 16.00% in four visuo-motor control tasks.

3533. Fine-Grained Verifiers: Preference Modeling as Next-token Prediction

in Vision-Language Alignment

链接：https://iclr.cc/virtual/2025/poster/29059 abstract：

3534. MMIE: Massive Multimodal Interleaved Comprehension Benchmark for Large Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/30211 abstract： Interleaved multimodal comprehension and generation, enabling models to produce and interpret both images and text in arbitrary sequences, have become a pivotal area in multimodal learning. Despite significant advancements, the evaluation of this capability remains insufficient. Existing benchmarks suffer from limitations in data scale, scope, and evaluation depth, while current evaluation metrics are often costly or biased, lacking in reliability for practical applications. To address these challenges, we introduce MMIE, a large-scale knowledge-intensive benchmark for evaluating interleaved multimodal comprehension and generation in Large Vision-Language Models (LVLMs). MMIE comprises 20K meticulously curated multimodal queries, spanning 3 categories, 12 fields, and 102 subfields, including mathematics, coding, physics, literature, health, and arts. It supports both interleaved inputs and outputs, offering a mix of multiple-choice and open-ended question formats to evaluate diverse competencies. Moreover, we propose a reliable automated evaluation metric, leveraging a scoring model fine-tuned with human-annotated data and systematic evaluation criteria, aimed at reducing bias and improving evaluation accuracy. Extensive experiments demonstrate the effectiveness of our benchmark and metrics in providing a comprehensive evaluation of interleaved LVLMs. Specifically, we evaluate eight LVLMs, revealing that even the best models show significant room for improvement, with most achieving only moderate results. We believe MMIE will drive further advancements in the development of interleaved LVLMs.

3535. Feast Your Eyes: Mixture-of-Resolution Adaptation for Multimodal Large Language Models

链接：https://iclr.cc/virtual/2025/poster/31217 abstract： In existing multimodal large language models (MLLMs), image resolution plays a significant role for granular visual recognition. However, directly increasing image resolution leads to expensive computational cost for MLLMs. In this paper, we reveal that a combination of low- and high-resolution visual features can efficiently mitigate this shortcoming. Based on this principle, we propose a novel and efficient method for MLLMs, termed Mixture-of-Resolution Adaptation (MRA). In particular, MRA adopts two visual pathways for images of different resolutions, where high-resolution visual information is embedded into the low-resolution pathway via the novel mixture-of-resolution adapters (MR-Adapters). This design also greatly reduces the input sequence length of MLLMs. To validate MRA, we apply it to a recent MLLM called LLaVA, and term the new model LLaVA-HR. We conduct extensive experiments on 17 vision-language (VL) tasks, which show that LLaVA-HR outperforms existing MLLMs on 15 VL tasks, e.g., +5.2\% on TextVQA. More importantly, both training and inference of LLaVA-HR remain efficient with MRA, e.g., 20 training hours and faster inference speed than LLaVA-NeXT. Source codes are released at: https://github.com/luogen1996/LLaVA-HR.

3536. Routing Experts: Learning to Route Dynamic Experts in Existing Multi-modal Large Language Models

链接：https://iclr.cc/virtual/2025/poster/27884 abstract： Recently, mixture of experts (MoE) has become a popular paradigm for achieving the trade-off between modal capacity and efficiency of multimodal large language models (MLLMs). Different from previous efforts, we are dedicated to exploring the dynamic experts in existing MLLMs and showing that a standard MLLM can also be a mixture of experts. However, achieving this target is still notoriously challenging. The well-trained MLLMs are more accustomed to the fixed pathway and a drastic change in its inference manner also greatly impedes its performance. To address these issues, we propose a novel dynamic expert routing method for existing MLLMs, termed Routing Experts (RoE), which can achieve example-dependent optimal path routing without obvious structure tweaks. Meanwhile, a new structure sparsity regularization is also introduced to force the well-trained MLLMs to learn more short-cut pathways. In addition, we also address the alignment of the training and inference of MLLMs in terms of network routing. To validate RoE, we apply it to a set of existing MLLMs, including LLaVA-1.5, LLaVA-HR and VILA, and conduct extensive experiments on a bunch of VL benchmarks. The experiment results not only show the effectiveness of our RoE in improving MLLMs' efficiency, but also yield obvious advantages over MoE-LLaVA in both performance and speed, e.g., an average performance gain of 3.3% on 5 benchmarks while being 1.61 times faster. Our code is anonymously released at https://github.com/DoubtedSteam/RoE

3537. $\gamma-$MoD: Exploring Mixture-of-Depth Adaptation for Multimodal Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28266 abstract： Despite the significant progress in multimodal large language models (MLLMs), their high computational cost remains a barrier to real-world deployment. Inspired by the mixture of depths (MoDs) in natural language processing, we aim to address this limitation from the perspective of ``activated tokens''. Our key insight is that if most tokens are redundant for the layer computation, then can be skipped directly via the MoD layer. However, directly converting the dense layers of MLLMs to MoD layers leads to substantial performance degradation. To address this issue, we propose an innovative MoD adaptation strategy for existing MLLMs called $\gamma$-MoD. In $\gamma$-MoD, a novel metric is proposed to guide the deployment of MoDs in the MLLM, namely rank of attention maps (ARank). Through ARank, we can effectively identify which layer is redundant and should be replaced with the MoD layer. Based on ARank, we further propose two novel designs to maximize the computational sparsity of MLLM while maintaining its performance, namely shared vision-language router and masked routing learning. With these designs, more than 90% dense layers of the MLLM can be effectively converted to the MoD ones. To validate our method, we apply it to three popular MLLMs, and conduct extensive experiments on 9 benchmark datasets. Experimental results not only validate the significant efficiency benefit of γ-MoD to existing MLLMs but also confirm its generalization ability on various MLLMs. For example, with a minor performance drop, i.e., -1.5%, γ-MoD can reduce the training and inference time of LLaVA-HR by 31.0% and 53.2%, respectively.

3538. DiffPC: Diffusion-based High Perceptual Fidelity Image Compression with Semantic Refinement

链接：https://iclr.cc/virtual/2025/poster/29654 abstract： Reconstructing high-quality images under low bitrates conditions presents a challenge, and previous methods have made this task feasible by leveraging the priors of diffusion models. However, the effective exploration of pre-trained latent diffusion models and semantic information integration in image compression tasks still needs further study. To address this issue, we introduce Diffusion-based High Perceptual Fidelity Image Compression with Semantic Refinement (DiffPC), a two-stage image compression framework based on stable diffusion. DiffPC efficiently encodes low-level image information, enabling the highly realistic reconstruction of the original image by leveraging high-level semantic features and the prior knowledge inherent in diffusion models. Specifically, DiffPC utilizes a multi-feature compressor to represent crucial low-level information with minimal bitrates and employs pre-embedding to acquire more robust hybrid semantics, thereby providing additional context for the decoding end. Furthermore, we have devised a control module tailored for image compression tasks, ensuring structural and textural consistency in reconstruction even at low bitrates and preventing decoding collapses induced by condition leakage. Extensive experiments demonstrate that our method achieves state-of-the-art perceptual fidelity and surpasses previous perceptual image compression methods by a significant margin in statistical fidelity.

3539. UniMatch: Universal Matching from Atom to Task for Few-Shot Drug Discovery

链接：https://iclr.cc/virtual/2025/poster/27929 abstract： Drug discovery is crucial for identifying candidate drugs for various diseases. However, its low success rate often results in a scarcity of annotations, posing a few-shot learning problem. Existing methods primarily focus on single-scale features, overlooking the hierarchical molecular structures that determine different molecular properties. To address these issues, we introduce Universal Matching Networks (UniMatch), a dual matching framework that integrates explicit hierarchical molecular matching with implicit task-level matching via meta-learning, bridging multi-level molecular representations and task-level generalization. Specifically, our approach explicitly captures structural features across multiple levels—atoms, substructures, and molecules—via hierarchical pooling and matching, facilitating precise molecular representation and comparison. Additionally, we employ a meta-learning strategy for implicit task-level matching, allowing the model to capture shared patterns across tasks and quickly adapt to new ones. This unified matching framework ensures effective molecular alignment while leveraging shared meta-knowledge for fast adaptation. Our experimental results demonstrate that UniMatch outperforms state-of-the-art methods on the MoleculeNet and FS-Mol benchmarks, achieving improvements of 2.87% in AUROC and 6.52% in ΔAUPRC. UniMatch also shows excellent generalization ability on the Meta-MolNet benchmark.

3540. Learning LLM-as-a-Judge for Preference Alignment

链接：https://iclr.cc/virtual/2025/poster/30221 abstract： Learning from preference feedback is a common practice for aligning large language models (LLMs) with human value. Conventionally, preference data is learned and encoded into a scalar reward model that connects a value head with an LLM to produce a scalar score as preference. However, scalar models lack interpretability and are known to be susceptible to biases in datasets. This paper investigates leveraging LLM itself to learn from such preference data and serve as a judge to address both limitations in one shot. Specifically, we prompt the pre-trained LLM to generate initial judgment pairs with contrastive preference in natural language form. The self-generated contrastive judgment pairs are used to train the LLM-as-a-Judge with Direct Preference Optimization (DPO) and incentivize its reasoning capability as a judge. This proposal of learning the LLMas-a-Judge using self-generated Contrastive judgments (Con-J) ensures natural interpretability through the generated rationales supporting the judgments, and demonstrates higher robustness against bias compared to scalar models. Experimental results show that Con-J outperforms the scalar reward model trained on the same collection of preference data, and outperforms a series of open-source and closed-source generative LLMs. We open-source the training process and model weights of Con-J at https://github.com/YeZiyi1998/Con-J.

3541. BitStack: Any-Size Compression of Large Language Models in Variable Memory Environments

链接：https://iclr.cc/virtual/2025/poster/28531 abstract：

3542. Data Mixing Laws: Optimizing Data Mixtures by Predicting Language

Modeling Performance

链接：https://iclr.cc/virtual/2025/poster/28624 abstract： Pretraining data of large language models composes multiple domains (e.g., web texts, academic papers, codes), whose mixture proportions crucially impact the competence of outcome models. While existing endeavors rely on heuristics or qualitative strategies to tune the proportions, we discover the quantitative predictability of model performance regarding the mixture proportions in function forms, which we refer to as the data mixing laws. Fitting such functions on sample mixtures unveils model performance on unseen mixtures before actual runs, thus guiding the selection of an ideal data mixture. Furthermore, we propose nested use of the scaling laws of training steps, model sizes, and our data mixing laws to predict the performance of large models trained on massive data under various mixtures with only small-scale training. Experimental results verify that our method effectively optimizes the training mixture of a 1B model trained for 100B tokens in RedPajama, reaching a performance comparable to the one trained for 48% more steps on the default mixture. Extending the application of data mixing laws to continual training accurately predicts the critical mixture proportion that avoids catastrophic forgetting and outlooks the potential for dynamic data schedules.

3543. Towards Universality: Studying Mechanistic Similarity Across Language Model Architectures

链接：https://iclr.cc/virtual/2025/poster/31148 abstract： The hypothesis of \textit{Universality} in interpretability suggests that different neural networks may converge toimplement similar algorithms on similar tasks. In this work, we investigate two mainstream architecturesfor language modeling, namely Transformers and Mambas, to explore the extent of their mechanistic similarity.We propose to use Sparse Autoencoders (SAEs) to isolate interpretable features from these models and showthat most features are similar in these two models. We also validate the correlation between feature similarityand~\univ. We then delve into the circuit-level analysis of Mamba modelsand find that the induction circuits in Mamba are structurally analogous to those in Transformers. We also identify a nuanced difference we call \emph{Off-by-One motif}: The information of one token is written into the SSM state in its next position. Whilst interaction between tokens in Transformers does not exhibit such trend.

3544. HQ-Edit: A High-Quality Dataset for Instruction-based Image Editing

链接：https://iclr.cc/virtual/2025/poster/32058 abstract： This study introduces HQ-Edit, a high-quality instruction-based image editing dataset with around 200,000 edits. Unlike prior approaches relying on attribute guidance or human feedback on building datasets, we devise a scalable data collection pipeline leveraging advanced foundation models, namely GPT-4V and DALL-E 3. To ensure its high quality, diverse examples are first collected online, expanded, and then used to create high-quality diptychs featuring input and output images with detailed text prompts, followed by precise alignment ensured through post-processing. In addition, we propose two evaluation metrics, Alignment and Coherence, to quantitatively assess the quality of image edit pairs using GPT-4V. HQ-Edits high-resolution images, rich in detail and accompanied by comprehensive editing prompts, substantially enhance the capabilities of existing image editing models. For example, an HQ-Edit finetuned InstructPix2Pix can attain state-of-the-art image editing performance, even surpassing those models fine-tuned with human-annotated data.

3545. MedTrinity-25M: A Large-scale Multimodal Dataset with Multigranular Annotations for Medicine

链接：https://iclr.cc/virtual/2025/poster/30141 abstract： This paper introduces MedTrinity-25M, a comprehensive, large-scale multimodal dataset for medicine, covering over 25 million images across 10 modalities with multigranular annotations for more than 65 diseases. These multigranular annotations encompass both global information, such as modality and organ detection, and local information like ROI analysis, lesion texture, and region-wise correlations. Unlike the existing multimodal datasets, which are limited by the availability of image-text pairs, we have developed the first automated pipeline that scales up multimodal data by generating multigranular visual and textual annotations in the form of image-ROI-description triplets without the need for any paired text descriptions. Specifically, data from over 30 different sources have been collected, preprocessed, and grounded using domain-specific expert models to identify ROIs related to abnormal regions. We then build a comprehensive knowledge base and prompt multimodal large language models to perform retrieval-augmented generation with the identified ROIs as guidance, resulting in multigranular textual descriptions. Compared to existing datasets, MedTrinity-25M provides the most enriched annotations, supporting a comprehensive range of multimodal tasks such as captioning and report generation, as well as vision-centric tasks like classification and segmentation. We propose LLaVA-Tri by pretraining LLaVA on MedTrinity-25M, achieving state-of-the-art performance on VQA-RAD, SLAKE, and PathVQA, surpassing representative SOTA multimodal large language models. Furthermore, MedTrinity-25M can also be utilized to support large-scale pre-training of multimodal medical AI models, contributing to the development of future foundation models in the medical domain. We will make our dataset available. The dataset is publicly available at https://yunfeixie233.github.io/MedTrinity-25M/.

3546. Reliable and Diverse Evaluation of LLM Medical Knowledge Mastery

链接：https://iclr.cc/virtual/2025/poster/29535 abstract： Mastering medical knowledge is crucial for medical-specific LLMs. However, despite the existence of medical benchmarks like MedQA, a unified framework that fully leverages existing knowledge bases to evaluate LLMs' mastery of medical knowledge is still lacking. We propose PretexEval, a novel framework that dynamically generates reliable and diverse test samples to evaluate LLMs for any given medical knowledge base. We notice that test samples produced directly from knowledge bases by templates or LLMs may introduce factual errors and also lack diversity. To address these issues, our framework employs predicate equivalence transformations to produce a series of variants for any given medical knowledge point. Finally, these produced predicate variants are converted into textual language, resulting in a series of reliable and diverse test samples. Here, we use our proposed framework to systematically investigate the mastery of medical factual knowledge of 12 well-known LLMs, based on two knowledge bases that are crucial for clinical diagnosis and treatment. The evaluation results illustrate that current LLMs still exhibit significant deficiencies in fully mastering medical knowledge, despite achieving considerable success on some famous public benchmarks. These new findings provide valuable insights for developing medical-specific LLMs, highlighting that current LLMs urgently need to strengthen their comprehensive and in-depth mastery of medical knowledge before being applied to real-world medical scenarios.

3547. On the Role of Attention Heads in Large Language Model Safety

链接：https://iclr.cc/virtual/2025/poster/28788 abstract： Large language models (LLMs) achieve state-of-the-art performance on multiple language tasks, yet their safety guardrails can be circumvented, leading to harmful generations. In light of this, recent research on safety mechanisms has emerged, revealing that when safety representations or component are suppressed, the safety capability of LLMs are compromised. However, existing research tends to overlook the safety impact of multi-head attention mechanisms, despite their crucial role in various model functionalities. Hence, in this paper, we aim to explore the connection between standard attention mechanisms and safety capability to fill this gap in the safety-related mechanistic interpretability. We propose an novel metric which tailored for multi-head attention, the Safety Head ImPortant Score (Ships), to assess the individual heads' contributions to model safety. Base on this, we generalize Ships to the dataset level and further introduce the Safety Attention Head AttRibution Algorithm (Sahara) to attribute the critical safety attention heads inside the model. Our findings show that special attention head has a significant impact on safety. Ablating a single safety head allows aligned model (e.g., Llama-2-7b-chat) to respond to **16$\times\uparrow$** more harmful queries, while only modifying **0.006\%** $\downarrow$ of the parameters, in contrast to the $\sim$ **5\%** modification required in previous studies. More importantly, we demonstrate that attention heads primarily function as feature extractors for safety and models fine-tuned from the same base model exhibit overlapping safety heads through comprehensive experiments. Together, our attribution approach and findings provide a novel perspective for unpacking the black box of safety mechanisms in large models.

3548. WeatherGFM: Learning a Weather Generalist Foundation Model via In-context Learning

链接：https://iclr.cc/virtual/2025/poster/28660 abstract： The Earth's weather system involves intricate weather data modalities and diverse weather understanding tasks, which hold significant value to human life. Existing data-driven models focus on single weather understanding tasks (e.g., weather forecasting). While these models have achieved promising results, they fail to tackle various complex tasks within a single and unified model. Moreover, the paradigm that relies on limited real observations for a single scenario hinders the model's performance upper bound.Inspired by the in-context learning paradigm from visual foundation models and large language models, in this paper, we introduce the first generalist weather generalist foundation model (WeatherGFM) to address weather understanding tasks in a unified manner. Specifically, we first unify the representation and definition for diverse weather understanding tasks.Subsequently, we design weather prompt formats to handle different weather data modalities, including single, multiple, and temporal modalities. Finally, we adopt a visual prompting question-answering paradigm for the training of unified weather understanding tasks. Extensive experiments indicate that our WeatherGFM can effectively handle up to 12 weather understanding tasks, including weather forecasting, super-resolution, weather image translation, and post-processing. Our method also showcases generalization ability on unseen tasks. The source code is available at https://github.com/xiangyu-mm/WeatherGFM.

3549. CARTS: Advancing Neural Theorem Proving with Diversified Tactic Calibration and Bias-Resistant Tree Search

链接：https://iclr.cc/virtual/2025/poster/29413 abstract： Recent advancements in neural theorem proving integrate large language models with tree search algorithms like Monte Carlo Tree Search (MCTS), where the language model suggests tactics and the tree search finds the complete proof path. However, many tactics proposed by the language model converge to semantically or strategically similar, reducing diversity and increasing search costs by expanding redundant proof paths. This issue exacerbates as computation scales and more tactics are explored per state. Furthermore, the trained value function suffers from false negatives, label imbalance, and domain gaps due to biased data construction. To address these challenges, we propose CARTS (diversified tactic CAlibration and bias-Resistant Tree Search), which balances tactic diversity and importance while calibrating model confidence. CARTS also introduce preference modeling and an adjustment term related to the ratio of valid tactics to improve the bias-resistance of the value function. Experimental results demonstrate that CARTS consistently outperforms previous methods achieving a pass@l rate of 49.6\% on the miniF2F-test benchmark. Further analysis confirms that CARTS improves tactic diversity and leads to a more balanced tree search.

3550. Circuit Representation Learning with Masked Gate Modeling and Verilog-AIG Alignment

链接：https://iclr.cc/virtual/2025/poster/29473 abstract： Understanding the structure and function of circuits is crucial for electronic design automation (EDA). Circuits can be formulated as And-Inverter graphs (AIGs), enabling efficient implementation of representation learning through graph neural networks (GNNs).Masked modeling paradigms have been proven effective in graph representation learning.However, masking augmentation to original circuits will destroy their logical equivalence, which is unsuitable for circuit representation learning.Moreover, existing masked modeling paradigms often prioritize structural information at the expense of abstract information such as circuit function.To address these limitations, we introduce MGVGA, a novel constrained masked modeling paradigm incorporating masked gate modeling (MGM) and Verilog-AIG alignment (VGA).Specifically, MGM preserves logical equivalence by masking gates in the latent space rather than in the original circuits, subsequently reconstructing the attributes of these masked gates.Meanwhile, large language models (LLMs) have demonstrated an excellent understanding of the Verilog code functionality.Building upon this capability, VGA performs masking operations on original circuits and reconstructs masked gates under the constraints of equivalent Verilog codes, enabling GNNs to learn circuit functions from LLMs.We evaluate MGVGA on various logic synthesis tasks for EDA and show the superior performance of MGVGA compared to previous state-of-the-art methods. Our code is available at https://github.com/wuhy68/MGVGA.

3551. The Lottery LLM Hypothesis, Rethinking What Abilities Should LLM Compression Preserve?

链接：https://iclr.cc/virtual/2025/poster/31348 abstract： Motivated by reducing the computational and storage costs of LLMs, model compression and KV cache compression have attracted much attention of researchers. However, Current methodologies predominantly emphasize maintaining the performance of compressed LLMs, as measured by perplexity or simple accuracy, on tasks involving common sense knowledge question answering and basic arithmetic reasoning. In this blog, we present a brief review of the recent advancements of LLM related to retrieval augmented generation, multi-step reasoning, external tools and computational expressivity, all of which substantially enhance LLM performance. Then, we propose a lottery LLM hypothesis suggesting that for a given LLM and task, there exists a smaller lottery LLM capable of producing the same performance with the original LLM with the assistances of multi-step reasoning and external tools. Based on the review of current progresses of LLMs, we discuss and summarize the essential capabilities that the lottery LLM and KV cache compression must possess, which are currently overlooked in existing methods.

3552. STBLLM: Breaking the 1-Bit Barrier with Structured Binary LLMs

链接：https://iclr.cc/virtual/2025/poster/30878 abstract： In this paper, we present the first structural binarization method for LLM compression to less than 1-bit precision. Although LLMs have achieved remarkable performance, their memory-bound nature during the inference stage hinders the adoption of resource-constrained devices. Reducing weights to 1-bit precision through binarization substantially enhances computational efficiency. We observe that randomly flipping some weights in binarized LLMs does not significantly degrade the model's performance, suggesting the potential for further compression. To exploit this, our STBLLM employs an N:M sparsity technique to achieve structural binarization of the weights. Specifically, we introduce a novel Standardized Importance (SI) metric, which considers weight magnitude and input feature norm to more accurately assess weight significance. Then, we propose a layer-wise approach, allowing different layers of the LLM to be sparsified with varying N:M ratios, thereby balancing compression and accuracy. Furthermore, we implement a fine-grained grouping strategy for less important weights, applying distinct quantization schemes to sparse, intermediate, and dense regions. Finally, we design a specialized CUDA kernel to support structural binarization. We conduct extensive experiments on LLaMA, OPT, and Mistral family. STBLLM achieves a perplexity of 11.07 at 0.55 bits per weight, outperforming the BiLLM by 3×. The results demonstrate that our approach performs better than other compressed binarization LLM methods while significantly reducing memory requirements. Code is released at https://github.com/pprp/STBLLM.

3553. Learning Graph Quantized Tokenizers

链接：https://iclr.cc/virtual/2025/poster/28351 abstract： Transformers serve as the backbone architectures of Foundational Models, where domain-specific tokenizers allow them to adapt to various domains. Graph Transformers (GTs) have recently emerged as leading models in geometric deep learning, outperforming Graph Neural Networks (GNNs) in various graph learning tasks. However, the development of tokenizers for graphs has lagged behind other modalities, with existing approaches relying on heuristics or GNNs co-trained with Transformers. To address this, we introduce GQT (\textbf{G}raph \textbf{Q}uantized \textbf{T}okenizer), which decouples tokenizer training from Transformer training by leveraging multi-task graph self-supervised learning, yielding robust and generalizable graph tokens. Furthermore, the GQT utilizes Residual Vector Quantization (RVQ) to learn hierarchical discrete tokens, resulting in significantly reduced memory requirements and improved generalization capabilities. By combining the GQT with token modulation, a Transformer encoder achieves state-of-the-art performance on 20 out of 22 benchmarks, including large-scale homophilic and heterophilic datasets. The implementation is publicly available at \href{https://github.com/limei0307/GQT}{https://github.com/limei0307/GQT}.

3554. Motion Control of High-Dimensional Musculoskeletal Systems with Hierarchical Model-Based Planning

链接：https://iclr.cc/virtual/2025/poster/29929 abstract： Controlling high-dimensional nonlinear systems, such as those found in biological and robotic applications, is challenging due to large state and action spaces. While deep reinforcement learning has achieved a number of successes in these domains, it is computationally intensive and time consuming, and therefore not suitable for solving large collections of tasks that require significant manual tuning. In this work, we introduce Model Predictive

Control with Morphology-aware Proportional Control (MPC^2), a hierarchical model-based learning algorithm for zero-shot and near-real-time control of high-dimensional complex dynamical systems. MPC^2 uses a sampling-based model predictive controller for target posture planning, and enables robust control for high-dimensional tasks by incorporating a morphology-aware proportional controller for actuator coordination. The algorithm enables motion control of a high-dimensional human musculoskeletal model in a variety of motion tasks, such as standing, walking on different terrains, and imitating sports activities. The reward function of MPC^2 can be tuned via black-box optimization, drastically reducing the need for human-intensive reward engineering.

3555. Stealthy Shield Defense: A Conditional Mutual Information-Based Approach against Black-Box Model Inversion Attacks

链接：https://iclr.cc/virtual/2025/poster/28325 abstract： Model inversion attacks (MIAs) aim to reconstruct the private training data by accessing a public model, raising concerns about privacy leakage. Black-box MIAs, where attackers can only query the model and obtain outputs, are closer to real-world scenarios. The latest black-box attacks have outperformed the state-of-the-art white-box attacks, and existing defenses cannot resist them effectively. To fill this gap, we propose Stealthy Shield Defense (SSD), a post-processing algorithm against black-box MIAs. Our idea is to modify the model's outputs to minimize the conditional mutual information (CMI). We mathematically prove that CMI is a special case of information bottlenecks (IB), and thus inherits the advantages of IB---making predictions less dependent on inputs and more dependent on ground truths. This theoretically guarantees our effectiveness, both in resisting MIAs and preserving utility. For minimizing CMI, we formulate a convex optimization problem and solve it via the water-filling method. Adaptive rate-distortion is introduced to constrain the modification to the outputs, and the water-filling is implemented on GPUs to address computation cost. Without the need to retrain the model, our algorithm is plug-and-play and easy to deploy. Experimental results indicate that SSD outperforms existing defenses, in terms of MIA resistance and model's utility, across various attack algorithms, training datasets, and model architectures. Our code is available at https://github.com/ZhuangQu/Stealthy-Shield-Defense.

3556. PathGen-1.6M: 1.6 Million Pathology Image-text Pairs Generation through Multi-agent Collaboration

链接：https://iclr.cc/virtual/2025/poster/28208 abstract： Vision Language Models (VLMs) like CLIP have attracted substantial attention in pathology, serving as backbones for applications such as zero-shot image classification and Whole Slide Image (WSI) analysis. Additionally, they can function as vision encoders when combined with large language models (LLMs) to support broader capabilities. Current efforts to train pathology VLMs rely on pathology image-text pairs from platforms like PubMed, YouTube, and Twitter, which provide limited, unscalable data with generally suboptimal image quality. In this work, we leverage large-scale WSI datasets like TCGA to extract numerous high-quality image patches. We then train a large multimodal model (LMM) to generate captions for extracted images, creating PathGen-1.6M, a dataset containing 1.6 million high-quality image-caption pairs. Our approach involves multiple agent models collaborating to extract representative WSI patches, generating and refining captions to obtain high-quality image-text pairs. Extensive experiments show that integrating these generated pairs with existing datasets to train a pathology-specific CLIP model, PathGen-CLIP, significantly enhances its ability to analyze pathological images, with substantial improvements across nine pathology-related zero-shot image classification tasks and three whole-slide image tasks. Furthermore, we construct 200K instruction-tuning data based on PathGen-1.6M and integrate PathGen-CLIP with the Vicuna LLM to create more powerful multimodal models through instruction tuning. Overall, we provide a scalable pathway for high-quality data generation in pathology, paving the way for next-generation general pathology models. Our dataset, code, and model are open-access at https://github.com/PathFoundation/PathGen-1.6M.

3557. BadRobot: Jailbreaking Embodied LLM Agents in the Physical World

链接：https://iclr.cc/virtual/2025/poster/32071 abstract： Embodied AI represents systems where AI is integrated into physical entities. Multimodal Large Language Model (LLM), which exhibits powerful language understanding abilities, has been extensively employed in embodied AI by facilitating sophisticated task planning. However, a critical safety issue remains overlooked: could these embodied LLMs perpetrate harmful behaviors? In response, we introduce BadRobot, the first attack paradigm designed to jailbreak robotic manipulation, making embodied LLMs violate safety and ethical constraints through typical voice-based user-system interactions. Specifically, three vulnerabilities are exploited to achieve this type of attack: (i) manipulation of LLMs within robotic systems, (ii) misalignment between linguistic outputs and physical actions, and (iii) unintentional hazardous behaviors caused by world knowledge's flaws. Furthermore, we construct a benchmark of various malicious physical action queries to evaluate BadRobot's attack performance. Based on this benchmark, extensive experiments against existing prominent embodied LLM frameworks (e.g., Voxposer, Code as Policies, and ProgPrompt) demonstrate the effectiveness of our BadRobot. We emphasize that addressing this emerging vulnerability is crucial for the secure deployment of LLMs in robotics.Warning: This paper contains harmful AI-generated language and aggressive actions.

3558. Hyper-Connections

链接：https://iclr.cc/virtual/2025/poster/30709 abstract： We present hyper-connections, a simple yet effective method that can serve as an alternative to residual connections. This approach specifically addresses common drawbacks observed in residual connection variants, such as the seesaw effect between gradient vanishing and representation collapse. Theoretically, hyper-connections allow the network to adjust the strength of connections between features at different depths and dynamically rearrange layers. We conduct experiments focusing on the pre-training of large language models, including dense and sparse models, where hyper-connections show significant performance improvements over residual connections. Additional experiments conducted on vision tasks also demonstrate similar improvements. We anticipate that this method will be broadly applicable and beneficial across a wide range of AI problems.

3559. Ultra-Sparse Memory Network

链接：https://iclr.cc/virtual/2025/poster/27653 abstract： It is widely acknowledged that the performance of Transformer models is logarithmically related to their number of parameters and computational complexity. While approaches like Mixture of Experts (MoE) decouple parameter count from computational complexity, they still face challenges in inference due to high memory access costs. This work introduces UltraMem, incorporating large-scale, ultra-sparse memory layer to address these limitations. Our approach significantly reduces inference latency while maintaining model performance. We also investigate the scaling laws of this new architecture, demonstrating that it not only exhibits favorable scaling properties but outperforms MoE. In experiments, the largest UltraMem we train has \textbf{20 million} memory slots. The results show that our method achieves state-of-the-art inference speed and model performance within a given computational budget, paving the way for billions of slots or experts.

3560. Re-Aligning Language to Visual Objects with an Agentic Workflow

链接：https://iclr.cc/virtual/2025/poster/29934 abstract： Language-based object detection (LOD) aims to align visual objects with language expressions. A large amount of paired data is utilized to improve LOD model generalizations. During the training process, recent studies leverage vision-language models (VLMs) to automatically generate human-like expressions for visual objects, facilitating training data scaling up. In this process, we observe that VLM hallucinations bring inaccurate object descriptions (e.g., object name, color, and shape) to deteriorate VL alignment quality. To reduce VLM hallucinations, we propose an agentic workflow controlled by an LLM to re-align language to visual objects via adaptively adjusting image and text prompts. We name this workflow Real-LOD, which includes planning, tool use, and reflection steps. Given an image with detected objects and VLM raw language expressions, Real-LOD reasons its state automatically and arranges action based on our neural symbolic designs (i.e., planning). The action will adaptively adjust the image and text prompts and send them to VLMs for object re-description (i.e., tool use). Then, we use another LLM to analyze these refined expressions for feedback (i.e., reflection). These steps are conducted in a cyclic form to gradually improve language descriptions for re-aligning to visual objects. We construct a dataset that contains a tiny amount of 0.18M images with re-aligned language expression and train a prevalent LOD model to surpass existing LOD methods by around 50% on the standard benchmarks. Our Real-LOD workflow, with automatic VL refinement, reveals a potential to preserve data quality along with scaling up data quantity, which further improves LOD performance from a data-alignment perspective.

3561. DSPO: Direct Score Preference Optimization for Diffusion Model Alignment

链接：https://iclr.cc/virtual/2025/poster/32046 abstract： Diffusion-based Text-to-Image (T2I) models have achieved impressive success in generating high-quality images from textual prompts. While large language models (LLMs) effectively leverage Direct Preference Optimization (DPO) for fine-tuning on human preference data without the need for reward models, diffusion models have not been extensively explored in this area. Current preference learning methods applied to T2I diffusion models immediately adapt existing techniques from LLMs. However, this direct adaptation introduces an estimated loss specific to T2I diffusion models. This estimation can potentially lead to suboptimal performance through our empirical results. In this work, we propose Direct Score Preference Optimization (DSPO), a novel algorithm that aligns the pretraining and fine-tuning objectives of diffusion models by leveraging score matching, the same objective used during pretraining. It introduces a new perspective on preference learning for diffusion models. Specifically, DSPO distills the score function of human-preferred image distributions into pretrained diffusion models, fine-tuning the model to generate outputs that align with human preferences. We theoretically show that DSPO shares the same optimization direction as reinforcement learning algorithms in diffusion models under certain conditions. Our experimental results demonstrate that DSPO outperforms preference learning baselines for T2I diffusion models in human preference evaluation tasks and enhances both visual appeal and prompt alignment of generated images.

3562. On Quantizing Neural Representation for Variable-Rate Video Coding

链接：https://iclr.cc/virtual/2025/poster/31039 abstract： This work introduces NeuroQuant, a novel post-training quantization (PTQ) approach tailored to non-generalized Implicit Neural Representations for variable-rate Video Coding (INR-VC). Unlike existing methods that require extensive weight retraining for each target bitrate, we hypothesize that variable-rate coding can be achieved by adjusting quantization parameters (QPs) of pre-trained weights. Our study reveals that traditional quantization methods, which assume inter-layer independence, are ineffective for non-generalized INR-VC models due to significant dependencies across layers. To address this, we redefine variable-rate INR-VC as a mixed-precision quantization problem and establish a theoretical framework for sensitivity criteria aimed at simplified, fine-grained rate control. Additionally, we propose network-wise calibration and channel-wise quantization strategies to minimize quantization-induced errors, arriving at a unified formula for representation-oriented PTQ calibration. Our experimental evaluations demonstrate that NeuroQuant significantly outperforms existing techniques in varying bitwidth quantization and compression efficiency, accelerating encoding by up to eight times and enabling quantization down to INT2 with minimal reconstruction loss. This work introduces variable-rate INR-VC for the first time and lays a theoretical foundation for future research in rate-distortion optimization, advancing the field of video coding technology. The materialswill be available at https://github.com/Eric-qi/NeuroQuant.

3563. From Decoupling to Adaptive Transformation: a Wider Optimization Space for PTQ

链接：https://iclr.cc/virtual/2025/poster/30116 abstract： Post-Training low-bit Quantization (PTQ) is useful to accelerate DNNs due to its high efficiency, the current SOTAs of which mostly adopt feature reconstruction with self-distillation finetuning. However, when bitwidth goes to be extremely low, we find the current reconstruction optimization space is not optimal. Considering all possible parameters and the ignored fact that integer weight can be obtained early before actual inference, we thoroughly explore different optimization space by quant-step decoupling, where a wider PTQ optimization space, which consistently makes a better optimum, is found out. Based on these, we propose an Adaptive Quantization Transformation (AdaQTransform) for PTQ reconstruction, which makes the quantized output feature better fit the FP32 counterpart with adaptive per-channel transformation, thus achieves lower feature reconstruction error. In addition, it incurs negligible extra finetuning cost and no extra inference cost. Based on AdaQTransform, for the first time, we build a general quantization setting paradigm subsuming current PTQs, QATs and other potential forms. Experiments demonstrate AdaQTransform expands the optimization space for PTQ and helps current PTQs find a better optimum over CNNs, ViTs, LLMs and image super-resolution networks, e.g., it improves NWQ by 5.7% on ImageNet for W2A2-MobileNet-v2.

3564. Towards Effective Evaluations and Comparisons for LLM Unlearning Methods

链接：https://iclr.cc/virtual/2025/poster/27842 abstract：

3565. Inverse Rendering using Multi-Bounce Path Tracing and Reservoir Sampling

链接：https://iclr.cc/virtual/2025/poster/30069 abstract： We introduce MIRReS, a novel two-stage inverse rendering framework thatjointly reconstructs and optimizes explicit geometry, materials, and lightingfrom multi-view images. Unlike previous methods that rely on implicit irradiance fields or oversimplified ray tracing, our method begins with an initialstage that extracts an explicit triangular mesh. In the second stage, we refine this representation using a physically-based inverse rendering modelwith multi-bounce path tracing and Monte Carlo integration. This enables our method to accurately estimate indirect illumination effects, including self-shadowing and internal reflections, leading to a more preciseintrinsic decomposition of shape, material, and lighting. To address thenoise issue in Monte Carlo integration, we incorporate reservoir sampling,improving convergence and enabling efficient gradient-based optimizationwith low sample counts. Through both qualitative and quantitative assessments across various scenarios, especially those with complex shadows,we demonstrate that our method achieves state-of-the-art decompositionperformance. Furthermore, our optimized explicit geometry seamlesslyintegrates with modern graphics engines supporting downstream applications such as scene editing, relighting, and material editing.

3566. An Intelligent Agentic System for Complex Image Restoration Problems

链接：https://iclr.cc/virtual/2025/poster/31076 abstract： Real-world image restoration (IR) is inherently complex and often requires combining multiple specialized models to address diverse degradations. Inspired by human problem-solving, we propose AgenticIR, an agentic system that mimics the human approach to image processing by following five key stages: Perception, Scheduling, Execution, Reflection, and Rescheduling. AgenticIR leverages large language models (LLMs) and vision-language models (VLMs) that interact via text generation to dynamically operate a toolbox of IR models. We fine-tune VLMs for image quality analysis and employ LLMs for reasoning, guiding the system step by step. To compensate for LLMs' lack of specific IR knowledge and experience, we introduce a self-exploration method, allowing the LLM to observe and summarize restoration results into referenceable documents. Experiments demonstrate AgenticIR's potential in handling complex IR tasks, representing a promising path toward achieving general intelligence in visual processing.

3567. Scaling Large Language Model-based Multi-Agent Collaboration

链接：https://iclr.cc/virtual/2025/poster/30077 abstract： Recent breakthroughs in large language model-driven autonomous agents have revealed that multi-agent collaboration often surpasses each individual through collective reasoning. Inspired by the neural scaling law—increasing neurons enhances performance, this study explores whether the continuous addition of collaborative agents can yield similar benefits. Technically, we utilize directed acyclic graphs to organize agents into a multi-agent collaboration network (MacNet), upon which their interactive reasoning is topologically orchestrated for autonomous task solving. Extensive evaluations reveal that it effectively supports collaboration among over a thousand agents, with irregular topologies outperforming regular ones. We also identify a collaborative scaling law—the overall performance follows a logistic growth pattern as agents scale, with collaborative emergence occurring earlier than traditional neural emergence. We speculate this may be because scaling agents catalyzes their multidimensional considerations during interactive reflection and refinement, thereby producing more comprehensive artifacts. The code is available at https://github.com/OpenBMB/ChatDev/tree/macnet.

3568. Toward Efficient Multi-Agent Exploration With Trajectory Entropy Maximization

链接：https://iclr.cc/virtual/2025/poster/29242 abstract： Recent works have increasingly focused on learning decentralized policies for agents as a solution to the scalability challenges in Multi-Agent Reinforcement Learning (MARL), where agents typically share the parameters of a policy network to make action decisions. However, this parameter sharing can impede efficient exploration, as it may lead to similar behaviors among agents. Different from previous mutual information-based methods that promote multi-agent diversity, we introduce a novel multi-agent exploration method called Trajectory Entropy Exploration (TEE). Our method employs a particle-based entropy estimator to maximize the entropy of different agents' trajectories in a contrastive trajectory representation space, resulting in diverse trajectories and efficient exploration. This entropy estimator avoids challenging density modeling and scales effectively in high-dimensional multi-agent settings. We integrate our method with MARL algorithms by deploying an intrinsic reward for each agent to encourage entropy maximization. To validate the effectiveness of our method, we test our method in challenging multi-agent tasks from several MARL benchmarks. The results demonstrate that our method consistently outperforms existing state-of-the-art methods.

3569. VideoGrain: Modulating Space-Time Attention for Multi-Grained Video Editing

链接：https://iclr.cc/virtual/2025/poster/29603 abstract： Recent advancements in diffusion models have significantly improved video generation and editing capabilities. However, multi-grained video editing, which encompasses class-level, instance-level, and part-level modifications, remains a formidable challenge. The major difficulties in multi-grained editing include semantic misalignment of text-to-region control and feature coupling within the diffusion model. To address these difficulties, we present VideoGrain, a zero-shot approach that modulates space-time (cross- and self-) attention mechanisms to achieve fine-grained control over video content. We enhance text-to-region control by amplifying each local prompt's attention to its corresponding spatial-disentangled region while minimizing interactions with irrelevant areas in cross-attention. Additionally, we improve feature separation by increasing intra-region awareness and reducing inter-region interference in self-attention. Extensive experiments demonstrate our method achieves state-of-the-art performance in real-world scenarios. Our code, data, and demos are available on the project page.

3570. Long-horizon Visual Instruction Generation with Logic and Attribute Self-reflection

链接：https://iclr.cc/virtual/2025/poster/30395 abstract： Visual instructions for long-horizon tasks are crucial as they intuitively clarify complex concepts and enhance retention across extended steps. Directly generating a series of images using text-to-image models without considering the context of previous steps results in inconsistent images, increasing cognitive load. Additionally, the generated images often miss objects or the attributes such as color, shape, and state of the objects are inaccurate.To address these challenges, we propose LIGER, the first training-free framework for Long-horizon Instruction GEneration with logic and attribute self-Reflection. LIGER first generates a draft image for each step with the historical prompt and visual memory of previous steps. This step-by-step generation approach maintains consistency between images in long-horizon tasks. Moreover, LIGER utilizes various image editing tools to rectify errors including wrong attributes, logic errors, object redundancy, and identity inconsistency in the draft images. Through this self-reflection mechanism, LIGER improves the logic and object attribute correctness of the images.To verify whether the generated images assist human understanding, we manually curated a new benchmark consisting of various long-horizon tasks. Human-annotated ground truth expressions reflect the human-defined criteria for how an image should appear to be illustrative. Experiments demonstrate the visual instructions generated by LIGER are more comprehensive compared with baseline methods. The code and dataset will be available once accepted.

3571. Less is More: Masking Elements in Image Condition Features Avoids Content Leakages in Style Transfer Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/32107 abstract： Given a style-reference image as the additional image condition, text-to-image diffusion models have demonstrated impressive capabilities in generating images that possess the content of text prompts while adopting the visual style of the reference image. However, current state-of-the-art methods often struggle to disentangle content and style from style-reference images, leading to issues such as content leakages. To address this issue, we propose a masking-based method that efficiently decouples content from style without the need of tuning any model parameters. By simply masking specific elements in the style reference's image features, we uncover a critical yet under-explored principle: guiding with appropriately-selected fewer conditions (e.g., dropping several image feature elements) can efficiently avoid unwanted content flowing into the diffusion models, enhancing the style transfer performances of text-to-image diffusion models. In this paper, we validate this finding both theoretically and experimentally. Extensive experiments across various styles demonstrate the effectiveness of our masking-based method and support our theoretical results.

3572. Noisy Test-Time Adaptation in Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/28661 abstract： Test-time adaptation (TTA) aims to address distribution shifts between source and target data by relying solely on target data during testing. In open-world scenarios, models often encounter noisy samples, i.e., samples outside the in-distribution (ID) label space. Leveraging the zero-shot capability of pre-trained vision-language models (VLMs), this paper introduces Zero-Shot Noisy TTA (ZS-NTTA), focusing on adapting the model to target data with noisy samples during test-time in a zero-shot manner. In the preliminary study, we reveal that existing TTA methods suffer from a severe performance decline under ZS-NTTA, often lagging behind even the frozen model. We conduct comprehensive experiments to analyze this phenomenon, revealing that the negative impact of unfiltered noisy data outweighs the benefits of clean data during model updating. In addition, as these methods adopt the adapting classifier to implement ID classification and noise detection sub-tasks, the ability of the model in both sub-tasks is largely hampered. Based on this analysis, we propose a novel framework that decouples the classifier and detector, focusing on developing an individual detector while keeping the classifier (including the backbone) frozen. Technically, we introduce the Adaptive Noise Detector (AdaND), which utilizes the frozen model's outputs as pseudo-labels to train a noise detector for detecting noisy samples effectively. To address clean data streams, we further inject Gaussian noise during adaptation, preventing the detector from misclassifying clean samples as noisy. Beyond the ZS-NTTA, AdaND can also improve the zero-shot out-of-distribution (ZS-OOD) detection ability of VLMs. Extensive experiments show that our method outperforms in both ZS-NTTA and ZS-OOD detection. On ImageNet, AdaND achieves a notable improvement of 8.32% in harmonic mean accuracy (Acc_H) for ZS-NTTA and 9.40% in FPR95 for ZS-OOD detection, compared to state-of-the-art methods. Importantly, AdaND is computationally efficient and comparable to the model-frozen method. The code is publicly available at: https://github.com/tmlr-group/ZS-NTTA.

3573. When Prompt Engineering Meets Software Engineering: CNL-P as Natural and Robust "APIs" for Human-AI Interaction

链接：https://iclr.cc/virtual/2025/poster/29577 abstract： With the growing capabilities of large language models (LLMs), they are increasingly applied in areas like intelligent customer service, code generation, and knowledge management. Natural language (NL) prompts act as the ``APIs'' for human-LLM interaction. To improve prompt quality, best practices for prompt engineering (PE) have been developed, including writing guidelines and templates. Building on this, we propose Controlled NL for Prompt (CNL-P), which not only incorporates PE best practices but also draws on key principles from software engineering (SE). CNL-P introduces precise grammar structures and strict semantic norms, further eliminating NL's ambiguity, allowing for a declarative but structured and accurate expression of user intent. This helps LLMs better interpret and execute the prompts, leading to more consistent and higher-quality outputs. We also introduce an NL2CNL-P conversion tool based on LLMs, enabling users to write prompts in NL, which are then transformed into CNL-P format, thus lowering the learning curve of CNL-P. In particular, we develop a linting tool that checks CNL-P prompts for syntactic and semantic accuracy, applying static analysis techniques to NL for the first time.Extensive experiments demonstrate that CNL-P enhances the quality of LLM responses through the novel and organic synergy of PE and SE. We believe that CNL-P can bridge the gap between emerging PE and traditional SE, laying the foundation for a new programming paradigm centered around NL.

3574. An Empirical Analysis of Uncertainty in Large Language Model Evaluations

链接：https://iclr.cc/virtual/2025/poster/30131 abstract： As LLM-as-a-Judge emerges as a new paradigm for assessing large language models (LLMs), concerns have been raised regarding the alignment, bias, and stability of LLM evaluators. While substantial work has focused on alignment and bias, little research has concentrated on the stability of LLM evaluators. In this paper, we conduct extensive experiments involving 9 widely used LLM evaluators across 2 different evaluation settings to investigate the uncertainty in model-based LLM evaluations. We pinpoint that LLM evaluators exhibit varying uncertainty based on model families and sizes. With careful comparative analyses, we find that employing special prompting strategies, whether during inference or post-training, can alleviate evaluation uncertainty to some extent. By utilizing uncertainty to enhance LLM's reliability and detection capability in Out-Of-Distribution (OOD) data, we further fine-tune an uncertainty-aware LLM evaluator named ConfiLM using a human-annotated fine-tuning set and assess ConfiLM's OOD evaluation ability on a manually designed test set sourced from the 2024 Olympics. Experimental results demonstrate that incorporating uncertainty as additional information during the fine-tuning phase can largely improve the model's evaluation performance in OOD scenarios. The code and data are released at: https://github.com/hasakiXie123/LLM-Evaluator-Uncertainty.

3575. Towards Out-of-Modal Generalization without Instance-level Modal Correspondence

链接：https://iclr.cc/virtual/2025/poster/29970 abstract： The world is understood from various modalities, such as appearance, sound, language, etc. Since each modality only partially represents objects in a certain physical meaning, leveraging additional ones is beneficial in both theory and practice. However, exploiting novel modalities normally requires cross-modal pairs corresponding to the same instance, which is extremely resource-consuming and sometimes even impossible, making knowledge exploration of novel modalities largely restricted. To seek practical multi-modal learning, here we study Out-of-Modal (OOM) Generalization as an initial attempt to generalize to an unknown modality without given instance-level modal correspondence. Specifically, we consider Semi-Supervised and Unsupervised scenarios of OOM Generalization, where the first has scarce correspondences and the second has none, and propose connect & explore (COX) to solve these problems. COX first connects OOM data and known In-Modal (IM) data through a variational information bottleneck framework to extract shared information. Then, COX leverages the shared knowledge to create emergent correspondences, which is theoretically justified from an information-theoretic perspective. As a result, the label information on OOM data emerges along with the correspondences, which help explore the OOM data with unknown knowledge, thus benefiting generalization results. We carefully evaluate the proposed COX method under various OOM generalization scenarios, verifying its effectiveness and extensibility.

3576. Quantifying Generalization Complexity for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28616 abstract： While large language models (LLMs) have shown exceptional capabilities in understanding complex queriesand performing sophisticated tasks, their generalization abilities are often deeply entangled with memorization, necessitating more precise evaluation.To address this challenge, we introduce Scylla, a dynamic evaluation framework that quantitatively measures the generalization abilities of LLMs. Scylla disentangles generalization from memorization via assessing model performance on both in-distribution (ID) and out-of-distribution (OOD) data through 20 tasks across 5 levels of complexity.Through extensive experiments, we uncover a non-monotonic relationship between task complexity and the performance gap between ID and OODdata, which we term the generalization valley.Specifically, this phenomenon reveals a critical threshold---referred toas critical complexity---where reliance on non-generalizable behavior peaks, indicating theupper bound of LLMs' generalization capabilities.As model size increases, the critical complexity shifts toward higher levels of task complexity, suggesting that larger models can handle more complex reasoning tasks before over-relying onmemorization.Leveraging Scylla and the concept of critical complexity, we benchmark 28 LLMs including both open-sourced models such as LLaMA and Qwen families, and closed-sourced models like Claude and GPT, providing a more robust evaluation and establishing a clearer understanding of LLMs' generalization capabilities.

3577. PixWizard: Versatile Image-to-Image Visual Assistant with Open-Language Instructions

链接：https://iclr.cc/virtual/2025/poster/32047 abstract： This paper presents a versatile image-to-image visual assistant, PixWizard, designed for image generation, manipulation, and translation based on free-from language instructions. To this end, we tackle a variety of vision tasks into a unified image-text-to-image generation framework and curate an Omni Pixel-to-Pixel Instruction-Tuning Dataset. By constructing detailed instruction templates in natural language, we comprehensively include a large set of diverse vision tasks such as text-to-image generation, image restoration, image grounding, dense image prediction, image editing, controllable generation, inpainting/outpainting, and more. Furthermore, we adopt Diffusion Transformers (DiT) as our foundation model and extend its capabilities with a flexible any resolution mechanism, enabling the model to dynamically process images based on the aspect ratio of the input, closely aligning with human perceptual processes. The model also incorporates structure-aware and semantic-aware guidance to facilitate effective fusion of information from the input image. Our experiments demonstrate that PixWizard not only shows impressive generative and understanding abilities for images with diverse resolutions but also exhibits generalization capabilities with unseen tasks and human instructions.

3578. LLaVA-MoD: Making LLaVA Tiny via MoE-Knowledge Distillation

链接：https://iclr.cc/virtual/2025/poster/27973 abstract： We introduce LLaVA-MoD, a novel framework designed to enable the efficient training of small-scale Multimodal Language Models ($s$-MLLM) distilling knowledge from large-scale MLLM ($l$-MLLM). Our approach tackles two fundamental challenges in MLLM distillation. First, we optimize the network structure of $s$-MLLM by integrating a sparse Mixture of Experts (MoE) architecture into the language model, striking a balance between computational efficiency and model expressiveness. Second, we propose a progressive knowledge transfer strategy for comprehensive knowledge transfer. This strategy begins with mimic distillation, where we minimize the Kullback-Leibler (KL) divergence between output distributions to enable $s$-MLLM to emulate $s$-MLLM's understanding. Following this, we introduce preference distillation via Preference Optimization (PO), where the key lies in treating $l$-MLLM as the reference model. During this phase, the $s$-MLLM's ability to discriminate between superior and inferior examples is significantly enhanced beyond $l$-MLLM, leading to a better $s$-MLLM that surpasses $l$-MLLM, particularly in hallucination benchmarks.Extensive experiments demonstrate that LLaVA-MoD surpasses existing works across various benchmarks while maintaining a minimal activated parameters and low computational costs. Remarkably, LLaVA-MoD-2B surpasses Qwen-VL-Chat-7B with an average gain of 8.8\%, using merely $0.3\%$ of the training data and 23\% trainable parameters. The results underscore LLaVA-MoD's ability to effectively distill comprehensive knowledge from its teacher model, paving the way for developing efficient MLLMs.

3579. AgentRefine: Enhancing Agent Generalization through Refinement Tuning

链接：https://iclr.cc/virtual/2025/poster/30355 abstract： Large Language Model (LLM) based agents have proved their ability to perform complex tasks like humans. However, there is still a large gap between open-sourced LLMs and commercial models like the GPT series. In this paper, we focus on improving the agent generalization capabilities of LLMs via instruction tuning. We first observe that the existing agent training corpus exhibits satisfactory results on held-in evaluation sets but fails to generalize to held-out sets. These agent-tuning works face severe formatting errors and are frequently stuck in the same mistake for a long while. We analyze that the poor generalization ability comes from overfitting to several manual agent environments and a lack of adaptation to new situations. They struggle with the wrong action steps and can not learn from the experience but just memorize existing observation-action relations. Inspired by the insight, we propose a novel AgentRefine framework for agent-tuning. The core idea is to enable the model to learn to correct its mistakes via observation in the trajectory. Specifically, we propose an agent synthesis framework to encompass a diverse array of environments and tasks and prompt a strong LLM to refine its error action according to the environment feedback. AgentRefine significantly outperforms state-of-the-art agent-tuning work in terms of generalization ability on diverse agent tasks. It also has better robustness facing perturbation and can generate diversified thought in inference. Our findings establish the correlation between agent generalization and self-refinement and provide a new paradigm for future research.

3580. CS-Bench: A Comprehensive Benchmark for Large Language Models towards Computer Science Mastery

链接：https://iclr.cc/virtual/2025/poster/28862 abstract： Large language models (LLMs) have demonstrated significant potential in advancing various fields of research and society. However, the current community of LLMs overly focuses on benchmarks for analyzing specific foundational skills (e.g. mathematics and code generation), neglecting an all-round evaluation of the computer science field. To bridge this gap, we introduce CS-Bench, the first multilingual (English, Chinese, French, German) benchmark dedicated to evaluating the performance of LLMs in computer science. CS-Bench comprises approximately 10K meticulously curated test samples, covering 26 subfields across 4 key areas of computer science, encompassing various task forms and divisions of knowledge and reasoning. Utilizing CS-Bench, we conduct a comprehensive evaluation of over 30 mainstream LLMs, revealing the relationship between CS performance and model scales. We also quantitatively analyze the reasons for failures in existing LLMs and highlight directions for improvements, including knowledge supplementation and CS-specific reasoning. Further cross-capability experiments show a high correlation between LLMs' capabilities in computer science and their abilities in mathematics and coding. Moreover, expert LLMs specialized in mathematics and coding also demonstrate strong performances in several CS subfields. Looking ahead, we envision CS-Bench serving as a cornerstone for LLM applications in the CS field and paving new avenues in assessing LLMs' diverse reasoning capabilities. Our project homepage is available at https://csbench.github.io/.

3581. MOS: Model Synergy for Test-Time Adaptation on LiDAR-Based 3D Object Detection

链接：https://iclr.cc/virtual/2025/poster/29272 abstract： LiDAR-based 3D object detection is crucial for various applications but often experiences performance degradation in real-world deployments due to domain shifts. While most studies focus on cross-dataset shifts, such as changes in environments and object geometries, practical corruptions from sensor variations and weather conditions remain underexplored. In this work, we propose a novel online test-time adaptation framework for 3D detectors that effectively tackles these shifts, including a challenging $\textit{cross-corruption}$ scenario where cross-dataset shifts and corruptions co-occur. By leveraging long-term knowledge from previous test batches, our approach mitigates catastrophic forgetting and adapts effectively to diverse shifts. Specifically, we propose a Model Synergy (MOS) strategy that dynamically selects historical checkpoints with diverse knowledge and assembles them to best accommodate the current test batch. This assembly is directed by our proposed Synergy Weights (SW), which perform a weighted averaging of the selected checkpoints, minimizing redundancy in the composite model. The SWs are computed by evaluating the similarity of predicted bounding boxes on the test data and the independence of features between checkpoint pairs in the model bank. To maintain an efficient and informative model bank, we discard checkpoints with the lowest average SW scores, replacing them with newly updated models. Our method was rigorously tested against existing test-time adaptation strategies across three datasets and eight types of corruptions, demonstrating superior adaptability to dynamic scenes and conditions. Notably, it achieved a 67.3% improvement in a challenging cross-corruption scenario, offering a more comprehensive benchmark for adaptation. Source code: https://github.com/zhuoxiao-chen/MOS.

3582. Vision Language Models are In-Context Value Learners

链接：https://iclr.cc/virtual/2025/poster/28853 abstract： Predicting temporal progress from visual trajectories is important for intelligent robots that can learn, adapt, and improve. However, learning such progress estimator, or temporal value function, across different tasks and domains requires both a large amount of diverse data and methods which can scale and generalize. To address these challenges, we present Generative Value Learning (GVL), a universal value function estimator that leverages the world knowledge embedded in vision-language models (VLMs) to predict task progress. Naively asking a VLM to predict values for a video sequence performs poorly due to the strong temporal correlation between successive frames. Instead, GVL poses value estimation as a temporal ordering problem over shuffled video frames; this seemingly more challenging task encourages VLMs to more fully exploit their underlying semantic and temporal grounding capabilities to differentiate frames based on their perceived task progress, consequently producing significantly better value predictions. Without any robot or task specific training, GVL can in-context zero-shot and few-shot predict effective values for more than 300 distinct real-world tasks across diverse robot platforms, including challenging bimanual manipulation tasks. Furthermore, we demonstrate that GVL permits flexible multi-modal in-context learning via examples from heterogeneous tasks and embodiments, such as human videos. The generality of GVL enables various downstream applications pertinent to visuomotor policy learning, including dataset filtering, success detection, and value-weighted regression -- all without any model training or finetuning.

3583. DynFrs: An Efficient Framework for Machine Unlearning in Random Forest

链接：https://iclr.cc/virtual/2025/poster/28390 abstract： Random Forests are widely recognized for establishing efficacy in classification and regression tasks, standing out in various domains such as medical diagnosis, finance, and personalized recommendations. These domains, however, are inherently sensitive to privacy concerns, as personal and confidential data are involved. With increasing demand for the right to be forgotten, particularly under regulations such as GDPR and CCPA, the ability to perform machine unlearning has become crucial for Random Forests. However, insufficient attention was paid to this topic, and existing approaches face difficulties in being applied to real-world scenarios. Addressing this gap, we propose the DynFrs framework designed to enable efficient machine unlearning in Random Forests while preserving predictive accuracy. Dynfrs leverages subsampling method Occ(q) and a lazy tag strategy Lzy, and is still adaptable to any Random Forest variant. In essence, Occ(q) ensures that each sample in the training set occurs only in a proportion of trees so that the impact of deleting samples is limited, and Lzy delays the reconstruction of a tree node until necessary, thereby avoiding unnecessary modifications on tree structures. In experiments, applying Dynfrs on Extremely Randomized Trees yields substantial improvements, achieving orders of magnitude faster unlearning performance and better predictive accuracy than existing machine unlearning methods for Random Forests.

3584. SigDiffusions: Score-Based Diffusion Models for Time Series via Log-Signature Embeddings

链接：https://iclr.cc/virtual/2025/poster/32077 abstract： Score-based diffusion models have recently emerged as state-of-the-art generativemodels for a variety of data modalities. Nonetheless, it remains unclear how toadapt these models to generate long multivariate time series. Viewing a timeseries as the discretisation of an underlying continuous process, we introduceSigDiffusion, a novel diffusion model operating on log-signature embeddingsof the data. The forward and backward processes gradually perturb and denoiselog-signatures while preserving their algebraic structure. To recover a signal fromits log-signature, we provide new closed-form inversion formulae expressing thecoefficients obtained by expanding the signal in a given basis (e.g. Fourier ororthogonal polynomials) as explicit polynomial functions of the log-signature.Finally, we show that combining SigDiffusions with these inversion formulaeresults in high-quality long time series generation, competitive with the currentstate-of-the-art on various datasets of synthetic and real-world examples.

3585. TEOChat: A Large Vision-Language Assistant for Temporal Earth Observation Data

链接：https://iclr.cc/virtual/2025/poster/28289 abstract：

3586. Asymmetric Factorized Bilinear Operation for Vision Transformer

链接：https://iclr.cc/virtual/2025/poster/29941 abstract： As a core component of Transformer-like deep architectures, a feed-forward network (FFN) for channel mixing is responsible for learning features of each token. Recent works show channel mixing can be enhanced by increasing computational burden or can be slimmed at the sacrifice of performance. Although some efforts have been made, existing works are still struggling to solve the paradox of performance and complexity trade-offs. In this paper, we propose an Asymmetric Factorized Bilinear Operation (AFBO) to replace FFN of vision transformer (ViT), which attempts to efficiently explore rich statistics of token features for achieving better performance and complexity trade-off. Specifically, our AFBO computes second-order statistics via a spatial-channel factorized bilinear operation for feature learning, which replaces a simple linear projection in FFN and enhances the feature learning ability of ViT by modeling second-order correlation among token features. Furthermore, our AFBO presents two structured-sparsity channel mapping strategies, namely Grouped Cross Channel Mapping (GCCM) and Overlapped Cycle Channel Mapping (OCCM). They decompose bilinear operation into grouped channel features by considering information interaction between groups, significantly reducing computational complexity while guaranteeing model performance. Finally, our AFBO is built with GCCM and OCCM in an asymmetric way, aiming to achieve a better trade-off. Note that our AFBO is model-agnostic, which can be flexibly integrated with existing ViTs. Experiments are conducted with twenty ViTs on various tasks, and the results show our AFBO is superior to its counterparts while improving existing ViTs in terms of generalization and robustness.

3587. Efficient Masked AutoEncoder for Video Object Counting and A Large-Scale Benchmark

链接：https://iclr.cc/virtual/2025/poster/28124 abstract： The dynamic imbalance of the fore-background is a major challenge in video object counting, which is usually caused by the sparsity of target objects. This remains understudied in existing works and often leads to severe under-/over-prediction errors. To tackle this issue in video object counting, we propose a density-embedded Efficient Masked Autoencoder Counting (E-MAC) framework in this paper. To empower the model's representation ability on density regression, we develop a new Density-Embedded Masked mOdeling (DEMO) method, which first takes the density map as an auxiliary modality to perform multimodal self-representation learning for image and density map. Although DEMO contributes to effective cross-modal regression guidance, it also brings in redundant background information, making it difficult to focus on the foreground regions. To handle this dilemma, we propose an efficient spatial adaptive masking derived from density maps to boost efficiency. Meanwhile, we employ an optical flow-based temporal collaborative fusion strategy to effectively capture the dynamic variations across frames, aligning features to derive multi-frame density residuals. The countingaccuracy of the current frame is boosted by harnessing the information from adjacent frames. In addition, considering that most existing datasets are limited to human-centric scenarios, we propose a large video bird counting dataset, DroneBird, in natural scenarios for migratory bird protection. Extensive experiments on three crowd datasets and our DroneBird validate our superiority againstthe counterparts. The code and dataset are available.

3588. DeepSeek-Prover-V1.5: Harnessing Proof Assistant Feedback for Reinforcement Learning and Monte-Carlo Tree Search

链接：https://iclr.cc/virtual/2025/poster/30193 abstract： Lean is an advanced proof assistant designed to facilitate formal theorem proving by providing a variety of interactive feedback. In this paper, we explore methodologies to leverage proof assistant feedback to augment the capabilities of large language models in constructing formal proofs. First, we deploy online reinforcement learning using Lean verification outcomes as the reward signal to improve the proof completion policy. This straightforward approach shows great promise in enhancing the model's alignment with the formal verification system. In addition, we propose RMaxTS, a variant of Monte-Carlo tree search that employs an intrinsic-reward-driven exploration strategy to generate diverse proof paths. The tree structure is organized to represent the transitions of intermediate tactic states, extracted from the compilation messages given by Lean's tactic mode. The intrinsic reward is constructed to incentivize the discovery of novel tactic states, which helps to to mitigate the sparse-reward problem inherent in proof search. These techniques lead to a more efficient planning scheme for formal proof generation, achieving new state-of-the-art results on both miniF2F and ProofNet benchmarks.

3589. DUET: Decentralized Bilevel Optimization without Lower-Level Strong Convexity

链接：https://iclr.cc/virtual/2025/poster/28607 abstract： Decentralized bilevel optimization (DBO) provides a powerful framework for multi-agent systems to solve local bilevel tasks in a decentralized fashion without the need for a central server. However, most existing DBO methods rely on lower-level strong convexity (LLSC) to guarantee unique solutions and a well-defined hypergradient for stationarity measure, hindering their applicability in many practical scenarios not satisfying LLSC. To overcome this limitation, we introduce a new single-loop DBO algorithm called diminishing quadratically-regularized bilevel decentralized optimization (DUET), which eliminates the need for LLSC by introducing a diminishing quadratic regularization to the lower-level (LL) objective. We show that DUET achieves an iteration complexity of $O(1/T^{1-5p-\frac{11}{4}\tau})$ for approximate KKT-stationary point convergence under relaxed assumptions, where p and $\tau $ are control parameters for LL learning rate and averaging, respectively.In addition, our DUET algorithm incorporates gradient tracking to address data heterogeneity, a key challenge in DBO settings. To the best of our knowledge, this is the first work to tackle DBO without LLSC under decentralized settings with data heterogeneity.Numerical experiments validate the theoretical findings and demonstrate the practical effectiveness of our proposed algorithms.

3590. Multi-Reward as Condition for Instruction-based Image Editing

链接：https://iclr.cc/virtual/2025/poster/30691 abstract： High-quality training triplets (instruction, original image, edited image) are essential for instruction-based image editing. Predominant training datasets (e.g., InsPix2Pix) are created using text-to-image generative models (e.g., Stable Diffusion, DALL-E) which are not trained for image editing. Accordingly, these datasets suffer from inaccurate instruction following, poor detail preserving, and generation artifacts. In this paper, we propose to address the training data quality issue with multi-perspective reward data instead of refining the ground-truth image quality. 1) we first design a quantitative metric system based on best-in-class LVLM (Large Vision Language Model), i.e., GPT-4o in our case, to evaluate the generation quality from 3 perspectives, namely, instruction following, detail preserving, and generation quality. For each perspective, we collected quantitative score in $0\sim 5$ and text descriptive feedback on the specific failure points in ground-truth edited images, resulting in a high-quality editing reward dataset, i.e., RewardEdit20K. 2) We further proposed a novel training framework to seamlessly integrate the metric output, regarded as multi-reward, into editing models to learn from the imperfect training triplets. During training, the reward scores and text descriptions are encoded as embeddings and fed into both the latent space and the U-Net of the editing models as auxiliary conditions. During inference, we set these additional conditions to the highest score with no text description for failure points, to aim at the best generation outcome. 3) We also build a challenging evaluation benchmark with real-world images/photos and diverse editing instructions, named as Real-Edit. Experiments indicate that our multi-reward conditioned model outperforms its no-reward counterpart on two popular editing pipelines, i.e., InsPix2Pix and SmartEdit. Code is released at https://github.com/bytedance/Multi-Reward-Editing.

3591. A Tight Convergence Analysis of Inexact Stochastic Proximal Point Algorithm for Stochastic Composite Optimization Problems

链接：https://iclr.cc/virtual/2025/poster/28431 abstract： The \textbf{i}nexact \textbf{s}tochastic \textbf{p}roximal \textbf{p}oint \textbf{a}lgorithm (isPPA) is popular for solving stochastic composite optimization problems with many applications in machine learning. While the convergence theory of the (inexact) PPA has been well established, the known convergence guarantees of isPPA require restrictive assumptions. In this paper, we establish the stability and almost sure convergence of isPPA under mild assumptions, where smoothness and (restrictive) strong convexity of the objective function are not required. Imposing a local Lipschitz condition on component functions and a quadratic growth condition on the objective function, we establish last-iterate iteration complexity bounds of isPPA regarding the distance to the solution set and the Karush–Kuhn–Tucker (KKT) residual.

Moreover, we show that the established iteration complexity bounds are tight up to a constant by explicitly analyzing the bounds for the regularized Fr\'echet mean problem. We further validate the established convergence guarantees of isPPA by numerical experiments.

3592. LongWriter: Unleashing 10,000+ Word Generation from Long Context LLMs

链接：https://iclr.cc/virtual/2025/poster/28586 abstract： Current long context large language models (LLMs) can process inputs up to 100,000 tokens, yet struggle to generate outputs exceeding even a modest length of 2,000 words. Through controlled experiments, we find that the model's effective generation length is inherently bounded by the sample it has seen during supervised fine-tuning (SFT). In other words, their output limitation is due to the scarcity of long-output examples in existing SFT datasets. To address this, we introduce AgentWrite, an agent-based pipeline that decomposes ultra-long generation tasks into subtasks, enabling off-the-shelf LLMs to generate coherent outputs exceeding 20,000 words. Leveraging AgentWrite, we construct LongWriter-6k, a dataset containing 6,000 SFT data with output lengths ranging from 2k to 32k words. By incorporating this dataset into model training, we successfully scale the output length of existing models to over 10,000 words while maintaining output quality. We also develop LongBench-Write, a comprehensive benchmark for evaluating ultra-long generation capabilities. Our 9B parameter model, further improved through DPO, achieves state-of-the-art performance on this benchmark, surpassing even much larger proprietary models. In general, our work demonstrates that existing long context LLM already possesses the potential for a larger output window--all you need is data with extended output during model alignment to unlock this capability.

3593. A Benchmark for Semantic Sensitive Information in LLMs Outputs

链接：https://iclr.cc/virtual/2025/poster/28322 abstract： Large language models (LLMs) can output sensitive information, which has emerged as a novel safety concern. Previous works focus on structured sensitive information (e.g. personal identifiable information). However, we notice that sensitive information can also be at semantic level, i.e. semantic sensitive information (SemSI). Particularly, simple natural questions can let state-of-the-art (SOTA) LLMs output SemSI. %which is hard to be detected compared with structured ones. Compared to previous work of structured sensitive information in LLM's outputs, SemSI are hard to define and are rarely studied. Therefore, we propose a novel and large-scale investigation on the existence of SemSI in SOTA LLMs induced by simple natural questions. First, we construct a comprehensive and labeled dataset of semantic sensitive information, SemSI-Set, by including three typical categories of SemSI. Then, we propose a large-scale benchmark, SemSI-Bench, to systematically evaluate semantic sensitive information in 25 SOTA LLMs. Our finding reveals that SemSI widely exists in SOTA LLMs' outputs by querying with simple natural questions.We open-source our project at https://semsi-project.github.io/.

3594. Evaluating Semantic Variation in Text-to-Image Synthesis: A Causal Perspective

链接：https://iclr.cc/virtual/2025/poster/29871 abstract： Accurate interpretation and visualization of human instructions are crucial for text-to-image (T2I) synthesis. However, current models struggle to capture semantic variations from word order changes, and existing evaluations, relying on indirect metrics like text-image similarity, fail to reliably assess these challenges. This often obscures poor performance on complex or uncommon linguistic patterns by the focus on frequent word combinations.To address these deficiencies, we propose a novel metric called SemVarEffect and a benchmark named SemVarBench, designed to evaluate the causality between semantic variations in inputs and outputs in T2I synthesis. Semantic variations are achieved through two types of linguistic permutations, while avoiding easily predictable literal variations.Experiments reveal that the CogView-3-Plus and Ideogram 2 performed the best, achieving a score of 0.2/1. Semantic variations in object relations are less understood than attributes, scoring 0.07/1 compared to 0.17-0.19/1. We found that cross-modal alignment in UNet or Transformers plays a crucial role in handling semantic variations, a factor previously overlooked by a focus on textual encoders. Our work establishes an effective evaluation framework that advances the T2I synthesis community's exploration of human instruction understanding. Our benchmark and code are available at https://github.com/zhuxiangru/SemVarBench.

3595. Data Center Cooling System Optimization Using Offline Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29382 abstract： The recent advances in information technology and artificial intelligence have fueled a rapid expansion of the data center (DC) industry worldwide, accompanied by an immense appetite for electricity to power the DCs. In a typical DC, around 30-40% of the energy is spent on the cooling system rather than on computer servers, posing a pressing need for developing new energy-saving optimization technologies for DC cooling systems. However, optimizing such real-world industrial systems faces numerous challenges, including but not limited to a lack of reliable simulation environments, limited historical data, and stringent safety and control robustness requirements. In this work, we present a novel physics-informed offline reinforcement learning (RL) framework for energy efficiency optimization of DC cooling systems. The proposed framework models the complex dynamical patterns and physical dependencies inside a server room using a purposely designed graph neural network architecture that is compliant with the fundamental time-reversal symmetry.

Because of its well-behaved and generalizable state-action representations, the model enables sample-efficient and robust latent space offline policy learning using limited real-world operational data. Our framework has been successfully deployed and verified in a large-scale production DC for closed-loop control of its air-cooling units (ACUs). We conducted a total of 2000 hours of short and long-term experiments in the production DC environment. The results show that our method achieves 14-21% energy savings in the DC cooling system, without any violation of the safety or operational constraints. We have also conducted a comprehensive evaluation of our approach in a real-world DC testbed environment. Our results have demonstrated the significant potential of offline RL in solving a broad range of data-limited, safety-critical real-world industrial control problems.

3596. ChartMimic: Evaluating LMM's Cross-Modal Reasoning Capability via Chart-to-Code Generation

链接：https://iclr.cc/virtual/2025/poster/28138 abstract： We introduce a new benchmark, ChartMimic, aimed at assessing the visually-grounded code generation capabilities of large multimodal models (LMMs). ChartMimic utilizes information-intensive visual charts and textual instructions as inputs, requiring LMMs to generate the corresponding code for chart rendering.ChartMimic includes $4,800$ human-curated (figure, instruction, code) triplets, which represent the authentic chart use cases found in scientific papers across various domains (e.g., Physics, Computer Science, Economics, etc). These charts span 18 regular types and 4 advanced types, diversifying into 201 subcategories.Furthermore, we propose multi-level evaluation metrics to provide an automatic and thorough assessment of the output code and the rendered charts.Unlike existing code generation benchmarks, ChartMimic places emphasis on evaluating LMMs' capacity to harmonize a blend of cognitive capabilities, encompassing visual understanding, code generation, and cross-modal reasoning. The evaluation of 3 proprietary models and 14 open-weight models highlights the substantial challenges posed by ChartMimic. Even the advanced GPT-4o, InternVL2-Llama3-76B only achieved an average score across Direct Mimic and Customized Mimic tasks of 82.2 and 61.6, respectively, indicating significant room for improvement. We anticipate that ChartMimic will inspire the development of LMMs, advancing the pursuit of artificial general intelligence.

3597. Lipschitz Bandits in Optimal Space

链接：https://iclr.cc/virtual/2025/poster/28717 abstract： This paper considers the Lipschitz bandit problem, where the set of arms is continuous and the expected reward is a Lipschitz function over the arm space. This problem has been extensively studied. Prior algorithms need to store the reward information of all visited arms, leading to significant memory consumption. We address this issue by introducing an algorithm named Log-space Lipschitz bandits (Log-Li), which achieves an optimal (up to logarithmic factors) regret of $\widetilde{O}\left(T^{\frac{d_z+1}{d_z+2}}\right)$ while only uses $O\left(\log T\right)$ bits of memory. Additionally, we provide a complexity analysis for this problem, demonstrating that $\Omega\left(\log T\right)$ bits of space are necessary for any algorithm to achieve the optimal regret. We also conduct numerical simulations, and the results show that our new algorithm achieves regret comparable to the state-of-the-art while reducing memory usage by orders of magnitude.

3598. TEASER: Token Enhanced Spatial Modeling for Expressions Reconstruction

链接：https://iclr.cc/virtual/2025/poster/28296 abstract： 3D facial reconstruction from a single in-the-wild image is a crucial task in human-centered computer vision tasks. While existing methods can recover accurate facial shapes, there remains significant space for improvement in fine-grained expression capture. Current approaches struggle with irregular mouth shapes, exaggerated expressions, and asymmetrical facial movements. We present TEASER (Token EnhAnced Spatial modeling for Expressions Reconstruction), which addresses these challenges and enhances 3D facial geometry performance. TEASER tackles two main limitations of existing methods: insufficient photometric loss for self-reconstruction and inaccurate localization of subtle expressions. We introduce a multi-scale tokenizer to extract facial appearance information. Combined with a neural renderer, these tokens provide precise geometric guidance for expression reconstruction. Furthermore, TEASER incorporates a pose-dependent landmark loss to further improve geometric performance. Our approach not only significantly enhances expression reconstruction quality but also offers interpretable tokens suitable for various downstream applications, such as photorealistic facial video driving, expression transfer, and identity swapping. Quantitative and qualitative experimental results across multiple datasets demonstrate that TEASER achieves state-of-the-art performance in precise expression reconstruction.

3599. INFER: A Neural-symbolic Model For Extrapolation Reasoning on Temporal Knowledge Graph

链接：https://iclr.cc/virtual/2025/poster/30378 abstract： Temporal Knowledge Graph(TKG) serves as an efficacious way to store dynamic facts in real-world. Extrapolation reasoning on TKGs, which aims at predicting possible future events, has attracted consistent research interest. Recently, some rule-based methods have been proposed, which are considered more interpretable compared with embedding-based methods. Existing rule-based methods apply rules through path matching or subgraph extraction, which falls short in inference ability and suffers from missing facts in TKGs. Besides, during rule application period, these methods consider the standing of facts as a binary 0 or 1 problem and ignores the validity as well as frequency of historical facts under temporal settings.In this paper, by designing a novel paradigm for rule application, we propose INFER, a neural-symbolic model for TKG extrapolation. With the introduction of Temporal Validity Function, INFER firstly considers the frequency and validity of historical facts and extends the truth value of facts into continuous real number to better adapt for temporal settings. INFER builds Temporal Weight Matrices with a pre-trained static KG embedding model to enhance its inference ability. Moreover, to facilitates potential integration with existing embedding-based methods, INFER adopts a rule projection module which enables it apply rules through conducting matrices operation on GPU. This feature also improves the efficiency of rule application. Experimental results show that INFER achieves state-of-the-art performance on various TKG datasets and significantly outperforms existing rule-based models on our modified, more sparse TKG datasets, which demonstrates the superiority of our model in inference ability.

3600. Generalizability of Neural Networks Minimizing Empirical Risk Based on Expressive Power

链接：https://iclr.cc/virtual/2025/poster/30734 abstract： The primary objective of learning methods is generalization. Classic generalization bounds, based on VC-dimension or Rademacher complexity, are uniformly applicable to all networks in the hypothesis space. On the other hand, algorithm-dependent generalization bounds, like stability bounds, address more practical scenarios and provide generalization conditions for neural networks trained using SGD. However, these bounds often rely on strict assumptions, such as the NTK hypothesis or convexity of the empirical loss, which are typically not met by neural networks. In order to establish generalizability under less stringent assumptions, this paper investigates generalizability of neural networks that minimize the empirical risk. A lower bound for population accuracy is established based on the expressiveness of these networks, which indicates that with adequately large training sample and network sizes, these networks can generalize effectively. Additionally, we provide a lower bound necessary for generalization, demonstrating that, for certain data distributions, the quantity of data required to ensure generalization exceeds the network size needed to represent that distribution. Finally, we provide theoretical insights into several phenomena in deep learning, including robust overfitting, importance of over-parameterization networks, and effects of loss functions.