201. Risk-Sensitive Variational Actor-Critic: A Model-Based Approach

链接：https://iclr.cc/virtual/2025/poster/28669 abstract： Risk-sensitive reinforcement learning (RL) with an entropic risk measure typically requires knowledge of the transition kernel or performs unstable updates w.r.t. exponential Bellman equations. As a consequence, algorithms that optimize this objective have been restricted to tabular or low-dimensional continuous environments. In this work we leverage the connection between the entropic risk measure and the RL-as-inference framework to develop a risk-sensitive variational actor-critic algorithm (rsVAC). Our work extends the variational framework to incorporate stochastic rewards and proposes a variational model-based actor-critic approach that modulates policy risk via a risk parameter. We consider, both, the risk-seeking and risk-averse regimes and present rsVAC learning variants for each setting. Our experiments demonstrate that this approach produces risk-sensitive policies and yields improvements in both tabular and risk-aware variants of complex continuous control tasks in MuJoCo.

202. Faster Diffusion Sampling with Randomized Midpoints: Sequential and Parallel

链接：https://iclr.cc/virtual/2025/poster/29931 abstract： Sampling algorithms play an important role in controlling the quality and runtime of diffusion model inference. In recent years, a number of works (Chen et al., 2023c;b; Benton et al., 2023; Lee et al., 2022) have analyzed algorithms for diffusion sampling with provable guarantees; these works show that for essentially any data distribution, one can approximately sample in polynomial time given a sufficiently accurate estimate of its score functions at different noise levels. In this work, we propose a new scheme inspired by Shen and Lee's randomized midpoint method for log-concave sampling (Shen & Lee, 2019). We prove that this approach achieves the best known dimension dependence for sampling from arbitrary smooth distributions in total variation distance ($\widetilde O(d^{5/12})$ compared to $\widetilde O(\sqrt{d})$ from prior work). We also show that our algorithm can be parallelized to run in only $\widetilde O(\log^2 d)$ parallel rounds, constituting the first provable guarantees for parallel sampling with diffusion models. As a byproduct of our methods, for the well-studied problem of log-concave sampling in total variation distance, we give an algorithm and simple analysis achieving dimension dependence $\widetilde O(d^{5/12})$ compared to $\widetilde O(\sqrt{d})$ from prior work.

203. Learning How Hard to Think: Input-Adaptive Allocation of LM Computation

链接：https://iclr.cc/virtual/2025/poster/30861 abstract： Computationally intensive decoding procedures---including search, reranking, and self-critique---can improve the quality of language model (LM) outputs in problems spanning code generation, numerical reasoning, and dialog.Existing work typically applies the same decoding procedure for every input to an LM. But not all inputs require the same amount of computation to process. Can we allocate decoding computation adaptively, using more resources to answer questions whose answers will be harder to compute? We present an approach that predicts the distribution of rewards given an input and computation budget, then allocates additional computation to inputs for which it is predicted to be most useful. We apply this approach in two decoding procedures: first, an adaptive best-of-k procedure that dynamically selects the number of samples to generate as input to a reranker; second, a routing procedure that dynamically responds to a query using a decoding procedure that is expensive but accurate, or one that is cheaper but less capable. Across a suite of programming, mathematics, and dialog tasks, we show that accurate computation-allocation procedures can be learned, and reduce computation by up to 50% at no cost to quality.

204. On Disentangled Training for Nonlinear Transform in Learned Image Compression

链接：https://iclr.cc/virtual/2025/poster/29489 abstract： Learned image compression (LIC) has demonstrated superior rate-distortion (R-D) performance compared to traditional codecs, but is challenged by training inefficiency that could incur more than two weeks to train a state-of-the-art model from scratch. Existing LIC methods overlook the slow convergence caused by compacting energy in learning nonlinear transforms. In this paper, we first reveal that such energy compaction consists of two components, \emph{i.e.}, feature decorrelation and uneven energy modulation. On such basis, we propose a linear auxiliary transform (AuxT) to disentangle energy compaction in training nonlinear transforms. The proposed AuxT obtains coarse approximation to achieve efficient energy compaction such that distribution fitting with the nonlinear transforms can be simplified to fine details. We then develop wavelet-based linear shortcuts (WLSs) for AuxT that leverages wavelet-based downsampling and orthogonal linear projection for feature decorrelation and subband-aware scaling for uneven energy modulation. AuxT is lightweight and plug-and-play to be integrated into diverse LIC models to address the slow convergence issue. Experimental results demonstrate that the proposed approach can accelerate training of LIC models by 2 times and simultaneously achieves an average 1\% BD-rate reduction. To our best knowledge, this is one of the first successful attempt that can significantly improve the convergence of LIC with comparable or superior rate-distortion performance.

205. GTR: Improving Large 3D Reconstruction Models through Geometry and Texture Refinement

链接：https://iclr.cc/virtual/2025/poster/29787 abstract： We propose a novel approach for 3D mesh reconstruction from multi-

view images. We improve upon the large reconstruction model LRM that use a transformer-based triplane generator and a Neural Radiance Field (NeRF) model trained on multi-view images. We introduce three key components to significantly enhance the 3D reconstruction quality. First of all, we examine the original LRM architecture and find several shortcomings. Subsequently, we introduce respective modifications to the LRM architecture, which lead to improved multi-view image representation and more computationally efficient training. Second, in order to improve geometry reconstruction and enable supervision at full image resolution, we extract meshes from the NeRF in a differentiable manner and fine-tune the NeRF model through mesh rendering. These modifications allow us to achieve state-of-the-art performance on both 2D and 3D evaluation metrics on Google Scanned Objects (GSO) dataset and OmniObject3D dataset. Finally, we introduce a lightweight per-instance texture refinement procedure to better reconstruct complex textures, such as text and portraits on assets. To address this, we introduce a lightweight per-instance texture refinement procedure. This procedure fine-tunes the triplane representation and the NeRF's color estimation model on the mesh surface using the input multi-view images in just 4 seconds. This refinement achieves faithful reconstruction of complex textures. Additionally, our approach enables various downstream applications, including text/image-to-3D generation.

206. VD3D: Taming Large Video Diffusion Transformers for 3D Camera Control

链接：https://iclr.cc/virtual/2025/poster/32114 abstract： Modern text-to-video synthesis models demonstrate coherent, photorealistic generation of complex videos from a text description. However, most existing models lack fine-grained control over camera movement, which is critical for downstream applications related to content creation, visual effects, and 3D vision. Recently, new methods demonstrate the ability to generate videos with controllable camera poses---these techniques leverage pre-trained U-Net-based diffusion models that explicitly disentangle spatial and temporal generation. Still, no existing approach enables camera control for new, transformer-based video diffusion models that process spatial and temporal information jointly. Here, we propose to tame video transformers for 3D camera control using a ControlNet-like conditioning mechanism that incorporates spatiotemporal camera embeddings based on Plucker coordinates. The approach demonstrates state-of-the-art performance for controllable video generation after fine-tuning on the RealEstate10K dataset. To the best of our knowledge, our work is the first to enable camera control for transformer-based video diffusion models.

207. Competition Dynamics Shape Algorithmic Phases of In-Context Learning

链接：https://iclr.cc/virtual/2025/poster/29296 abstract： In-Context Learning (ICL) has significantly expanded the general-purpose nature of large language models, allowing them to adapt to novel tasks using merely the inputted context. This has motivated a series of papers that analyze tractable synthetic domains and postulate precise mechanisms that may underlie ICL. However, the use of relatively distinct setups that often lack a sequence modeling nature to them makes it unclear how general the reported insights from such studies are. Motivated by this, we propose a synthetic sequence modeling task that involves learning to simulate a finite mixture of Markov chains. As we show, models trained on this task reproduce most well-known results on ICL, hence offering a unified setting for studying the concept. Building on this setup, we demonstrate we can explain a model’s behavior by decomposing it into four broad algorithms that combine a fuzzy retrieval vs. inference approach with either unigram or bigram statistics of the context. These algorithms engage in a competitive dynamics to dominate model behavior, with the precise experimental conditions dictating which algorithm ends up superseding others: e.g., we find merely varying context size or amount of training yields (at times sharp) transitions between which algorithm dictates the model behavior, revealing a mechanism that explains the transient nature of ICL. In this sense, we argue ICL is best thought of as a mixture of different algorithms, each with its own peculiarities, instead of a monolithic capability. This also implies that making general claims about ICL that hold universally across all settings may be infeasible.

208. ADAM Optimization with Adaptive Batch Selection

链接：https://iclr.cc/virtual/2025/poster/30565 abstract： Adam is a widely used optimizer in neural network training due to its adaptive learning rate. However, because different data samples influence model updates to varying degrees, treating them equally can lead to inefficient convergence. To address this, a prior work proposed adapting the sampling distribution using a bandit framework to select samples adaptively. While promising, both the original Adam and its bandit-based variant suffer from flawed theoretical guarantees. In this paper, we introduce Adam with Combinatorial Bandit Sampling (AdamCB), which integrates combinatorial bandit techniques into Adam to resolve these issues. AdamCB is able to fully utilize feedback from multiple actions at once, enhancing both theoretical guarantees and practical performance. Our rigorous regret analysis shows that AdamCB achieves faster convergence than both the original Adam and its variants. Numerical experiments demonstrate that AdamCB consistently outperforms existing Adam-based methods, making it the first to offer both provable guarantees and practical efficiency for Adam with adaptive batch selection.

209. Navigating Neural Space: Revisiting Concept Activation Vectors to Overcome Directional Divergence

链接：https://iclr.cc/virtual/2025/poster/29707 abstract： With a growing interest in understanding neural network prediction strategies, Concept Activation Vectors (CAVs) have emerged as a popular tool for modeling human-understandable concepts in

the latent space.Commonly, CAVs are computed by leveraging linear classifiers optimizing the separability of latent representations of samples with and without a given concept. However, in this paper we show that such a separability-oriented computation leads to solutions, which may diverge from the actual goal of precisely modeling the concept direction.This discrepancy can be attributed to the significant influence of distractor directions, i.e., signals unrelated to the concept, which are picked up by filters (i.e., weights) of linear models to optimize class-separability.To address this, we introduce pattern-based CAVs, solely focussing on concept signals, thereby providing more accurate concept directions.We evaluate various CAV methods in terms of their alignment with the true concept direction and their impact on CAV applications, including concept sensitivity testing and model correction for shortcut behavior caused by data artifacts. We demonstrate the benefits of pattern-based CAVs using the Pediatric Bone Age, ISIC2019, and FunnyBirds datasets with VGG, ResNet, ReXNet, EfficientNet, and Vision Transformer as model architectures.

210. Towards Realistic UAV Vision-Language Navigation: Platform, Benchmark, and Methodology

链接：https://iclr.cc/virtual/2025/poster/28193 abstract： Developing agents capable of navigating to a target location based on language instructions and visual information, known as vision-language navigation (VLN), has attracted widespread interest. Most research has focused on ground-based agents, while UAV-based VLN remains relatively underexplored. Recent efforts in UAV vision-language navigation predominantly adopt ground-based VLN settings, relying on predefined discrete action spaces and neglecting the inherent disparities in agent movement dynamics and the complexity of navigation tasks between ground and aerial environments. To address these disparities and challenges, we propose solutions from three perspectives: platform, benchmark, and methodology. To enable realistic UAV trajectory simulation in VLN tasks, we propose the OpenUAV platform, which features diverse environments, realistic flight control, and extensive algorithmic support. We further construct a target-oriented VLN dataset consisting of approximately 12k trajectories on this platform, serving as the first dataset specifically designed for realistic UAV VLN tasks. To tackle the challenges posed by complex aerial environments, we propose an assistant-guided UAV object search benchmark called UAV-Need-Help, which provides varying levels of guidance information to help UAVs better accomplish realistic VLN tasks. We also propose a UAV navigation LLM that, given multi-view images, task descriptions, and assistant instructions, leverages the multimodal understanding capabilities of the MLLM to jointly process visual and textual information, and performs hierarchical trajectory generation. The evaluation results of our method significantly outperform the baseline models, while there remains a considerable gap between our results and those achieved by human operators, underscoring the challenge presented by the UAV-Need-Help task.

211. Unsupervised Model Tree Heritage Recovery

链接：https://iclr.cc/virtual/2025/poster/29687 abstract： The number of models shared online has recently skyrocketed, with over one million public models available on Hugging Face. Sharing models allows other users to build on existing models, using them as initialization for fine-tuning, improving accuracy, and saving compute and energy. However, it also raises important intellectual property issues, as fine-tuning may violate the license terms of the original model or that of its training data. A Model Tree, i.e., a tree data structure rooted at a foundation model and having directed edges between a parent model and other models directly fine-tuned from it (children), would settle such disputes by making the model heritage explicit. Unfortunately, current models are not well documented, with most model metadata (e.g., "model cards") not providing accurate information about heritage. In this paper, we introduce the task of Unsupervised Model Tree Heritage Recovery (Unsupervised MoTHer Recovery) for collections of neural networks. For each pair of models, this task requires: i) determining if they are directly related, and ii) establishing the direction of the relationship. Our hypothesis is that model weights encode this information, the challenge is to decode the underlying tree structure given the weights. We discover several properties of model weights that allow us to perform this task. By using these properties, we formulate the MoTHer Recovery task as finding a directed minimal spanning tree. In extensive experiments we demonstrate that our method successfully reconstructs complex Model Trees.

212. Optimality of Matrix Mechanism on ℓ_p^p-metric

链接：https://iclr.cc/virtual/2025/poster/28869 abstract： In this paper, we introduce the ℓ_p^p-error metric (for $p \geq 2$) when answering linear queries under the constraint of differential privacy. We characterize such an error under (ϵ,δ)-differential privacy in the natural add/remove model. Before this paper, tight characterization in the hardness of privately answering linear queries was known under ℓ_2^2-error metric (Edmonds et al. 2020) and ℓ_p^2-error metric for unbiased mechanisms in the substitution model (Nikolov et al. 2024). As a direct consequence of our results, we give tight bounds on answering prefix sum and parity queries under differential privacy for all constant p in terms of the ℓ_p^p error, generalizing the bounds in Hhenzinger et al. for $p=2$.

213. PINP: Physics-Informed Neural Predictor with latent estimation of fluid flows

链接：https://iclr.cc/virtual/2025/poster/27928 abstract： Accurately predicting fluid dynamics and evolution has been a long-standing challenge in physical sciences. Conventional deep learning methods often rely on the nonlinear modeling capabilities of neural networks to establish mappings between past and future states, overlooking the fluid dynamics, or only modeling the velocity field, neglecting the coupling of multiple physical quantities. In this paper, we propose a new physics-informed learning approach that incorporates coupled physical quantities into the prediction process to assist with forecasting. Central to our

method lies in the discretization of physical equations, which are directly integrated into the model architecture and loss function. This integration enables the model to provide robust, long-term future predictions. By incorporating physical equations, our model demonstrates temporal extrapolation and spatial generalization capabilities. Experimental results show that our approach achieves the state-of-the-art performance in spatiotemporal prediction across both numerical simulations and real-world extreme-precipitation nowcasting benchmarks.

214. Adversarial Mixup Unlearning

链接：https://iclr.cc/virtual/2025/poster/30273 abstract： Machine unlearning is a critical area of research aimed at safeguarding data privacy by enabling the removal of sensitive information from machine learning models. One unique challenge in this field is catastrophic unlearning, where erasing specific data from a well-trained model unintentionally removes essential knowledge, causing the model to deviate significantly from a retrained one. To address this, we introduce a novel approach that regularizes the unlearning process by utilizing synthesized mixup samples, which simulate the data susceptible to catastrophic effects. At the core of our approach is a generator-unlearner framework, MixUnlearn, where a generator adversarially produces challenging mixup examples, and the unlearner effectively forgets target information based on these synthesized data. Specifically, we first introduce a novel contrastive objective to train the generator in an adversarial direction: generating examples that prompt the unlearner to reveal information that should be forgotten, while losing essential knowledge. Then the unlearner, guided by two other contrastive loss terms, processes the synthesized and real data jointly to ensure accurate unlearning without losing critical knowledge, overcoming catastrophic effects. Extensive evaluations across benchmark datasets demonstrate that our method significantly outperforms state-of-the-art approaches, offering a robust solution to machine unlearning. This work not only deepens understanding of unlearning mechanisms but also lays the foundation for effective machine unlearning with mixup augmentation.

215. Pushing the Limits of All-Atom Geometric Graph Neural Networks: Pre-Training, Scaling, and Zero-Shot Transfer

链接：https://iclr.cc/virtual/2025/poster/31010 abstract： The ability to construct transferable descriptors for molecular and biological systems has broad applications in drug discovery, molecular dynamics, and protein analysis. Geometric graph neural networks (Geom-GNNs) utilizing all-atom information have revolutionized atomistic simulations by enabling the prediction of interatomic potentials and molecular properties. Despite these advances, the application of all-atom Geom-GNNs in protein modeling remains limited due to computational constraints. In this work, we first demonstrate the potential of pre-trained Geom-GNNs as zero-shot transfer learners, effectively modeling protein systems with all-atom granularity. Through extensive experimentation to evaluate their expressive power, we characterize the scaling behaviors of Geom-GNNs across self-supervised, supervised, and unsupervised setups. Interestingly, we find that Geom-GNNs deviate from conventional power-law scaling observed in other domains, with no predictable scaling principles for molecular representation learning. Furthermore, we show how pre-trained graph embeddings can be directly used for analysis and synergize with other architectures to enhance expressive power for protein modeling.

216. Time-to-Event Pretraining for 3D Medical Imaging

链接：https://iclr.cc/virtual/2025/poster/27661 abstract： With the rise of medical foundation models and the growing availability of imaging data, scalable pretraining techniques offer a promising way to identify imaging biomarkers predictive of future disease risk. While current self-supervised methods for 3D medical imaging models capture local structural features like organ morphology, they fail to link pixel biomarkers with long-term health outcomes due to a missing context problem. Current approaches lack the temporal context necessary to identify biomarkers correlated with disease progression, as they rely on supervision derived only from images and concurrent text descriptions. To address this, we introduce time-to-event pretraining, a pretraining framework for 3D medical imaging models that leverages large-scale temporal supervision from paired, longitudinal electronic health records (EHRs). Using a dataset of 18,945 CT scans (4.2 million 2D images) and time-to-event distributions across thousands of EHR-derived tasks, our method improves outcome prediction, achieving an average AUROC increase of 23.7% and a 29.4% gain in Harrell's C-index across 8 benchmark tasks. Importantly, these gains are achieved without sacrificing diagnostic classification performance. This study lays the foundation for integrating longitudinal EHR and 3D imaging data to advance clinical risk prediction.

217. Real-Time Video Generation with Pyramid Attention Broadcast

链接：https://iclr.cc/virtual/2025/poster/28773 abstract： We present Pyramid Attention Broadcast (PAB), a real-time, high quality and training-free approach for DiT-based video generation. Our method is founded on the observation that attention difference in the diffusion process exhibits a U-shaped pattern, indicating significant redundancy. We mitigate this by broadcasting attention outputs to subsequent steps in a pyramid style. It applies different broadcast strategies to each attention based on their variance for best efficiency. We further introduce broadcast sequence parallel for more efficient distributed inference. PAB demonstrates up to 10.5x speedup across three models compared to baselines, achieving real-time generation for up to 720p videos. We anticipate that our simple yet effective method will serve as a robust baseline and facilitate future research and application for video generation.

218. SBSC: Step-by-Step Coding for Improving Mathematical Olympiad

Performance

链接：https://iclr.cc/virtual/2025/poster/27845 abstract： We propose Step-by-Step Coding (SBSC): a multi-turn math reasoning framework that enables Large Language Models (LLMs) to generate sequence of programs for solving Olympiad level math problems. After each turn/step, by leveraging the code execution outputs and programs of previous steps, the model generates the next sub-task and the corresponding program to complete it. This way, SBSC, sequentially navigates to reach the final answer. SBSC allows more granular, flexible and precise approach to problem-solving compared to existing methods. Extensive experiments highlight the effectiveness of SBSC in tackling competition and Olympiad-level math problems. For Claude-3.5-Sonnet, we observe SBSC (greedy decoding) surpasses existing state-of-the-art (SOTA) program generation based reasoning strategies by absolute 10.7% on AMC12, 8% on AIME and 12.6% on MathOdyssey. Given SBSC is multi-turn in nature, we also benchmark SBSC’s greedy decoding against self- consistency decoding results of existing SOTA math reasoning strategies and observe performance gain by absolute 6.2% on AMC, 6.7% on AIME and 7.4% on MathOdyssey.

219. Denoising Autoregressive Transformers for Scalable Text-to-Image Generation

链接：https://iclr.cc/virtual/2025/poster/29151 abstract： Diffusion models have become the dominant approach for visual generation. They are trained by denoising a Markovian process which gradually adds noise to the input. We argue that the Markovian property limits the model’s ability to fully utilize the generation trajectory, leading to inefficiencies during training and inference. In this paper, we propose DART, a transformer-based model that unifies autoregressive (AR) and diffusion within a non-Markovian framework. DART iteratively denoises image patches spatially and spectrally using an AR model that has the same architecture as standard language models. DART does not rely on image quantization, which enables more effective image modeling while maintaining flexibility. Furthermore, DART seamlessly trains with both text and image data in a unified model. Our approach demonstrates competitive performance on class-conditioned and text-to-image generation tasks, offering a scalable, efficient alternative to traditional diffusion models. Through this unified framework, DART sets a new benchmark for scalable, high-quality image synthesis.

220. Few-Class Arena: A Benchmark for Efficient Selection of Vision Models and Dataset Difficulty Measurement

链接：https://iclr.cc/virtual/2025/poster/31153 abstract： We propose Few-Class Arena (FCA), as a unified benchmark with focus on testing efficient image classification models for few classes. A wide variety of benchmark datasets with many classes (80-1000) have been created to assist Computer Vision architectural evolution. An increasing number of vision models are evaluated with these many-class datasets. However, real-world applications often involve substantially fewer classes of interest (2-10). This gap between many and few classes makes it difficult to predict performance of the few-class applications using models trained on the available many-class datasets. To date, little has been offered to evaluate models in this Few-Class Regime. We conduct a systematic evaluation of the ResNet family trained on ImageNet subsets from 2 to 1000 classes, and test a wide spectrum of Convolutional Neural Networks and Transformer architectures over ten datasets by using our newly proposed FCA tool. Furthermore, to aid an up-front assessment of dataset difficulty and a more efficient selection of models, we incorporate a difficulty measure as a function of class similarity. FCA offers a new tool for efficient machine learning in the Few-Class Regime, with goals ranging from a new efficient class similarity proposal, to lightweight model architecture design, to a new scaling law. FCA is user-friendly and can be easily extended to new models and datasets, facilitating future research work. Our benchmark is available at https://github.com/bryanbocao/fca.

221. GSBA^K: top-K Geometric Score-based Black-box Attack

链接：https://iclr.cc/virtual/2025/poster/28735 abstract： Existing score-based adversarial attacks mainly focus on crafting top-1 adversarial examples against classifiers with single-label classification. Their attack success rate and query efficiency are often less than satisfactory, particularly under small perturbation requirements; moreover, the vulnerability of classifiers with multi-label learning is yet to be studied. In this paper, we propose a comprehensive surrogate free score-based attack, named \b geometric \b score-based \b black-box \b attack (GSBA^K), to craft adversarial examples in an aggressive top-K setting for both untargeted and targeted attacks, where the goal is to change the top-K predictions of the target classifier. We introduce novel gradient-based methods to find a good initial boundary point to attack. Our iterative method employs novel gradient estimation techniques, particularly effective in top-K setting, on the decision boundary to effectively exploit the geometry of the decision boundary. Additionally, GSBA^K can be used to attack against classifiers with top-K multi-label learning. Extensive experiential results on ImageNet and PASCAL VOC datasets validate the effectiveness of GSBA^K in crafting top-K adversarial examples.

222. Logicbreaks: A Framework for Understanding Subversion of Rule-based Inference

链接：https://iclr.cc/virtual/2025/poster/28281 abstract： We study how to subvert large language models (LLMs) from following prompt-specified rules.We first formalize rule-following as inference in propositional Horn logic, a mathematical system in which rules have the form "if P and Q, then R" for some propositions P, Q, and R.Next, we prove that although small

transformers can faithfully follow such rules, maliciously crafted prompts can still mislead both theoretical constructions and models learned from data.Furthermore, we demonstrate that popular attack algorithms on LLMs find adversarial prompts and induce attention patterns that align with our theory.Our novel logic-based framework provides a foundation for studying LLMs in rule-based settings, enabling a formal analysis of tasks like logical reasoning and jailbreak attacks.

223. Bad-PFL: Exploiting Backdoor Attacks against Personalized Federated Learning

链接：https://iclr.cc/virtual/2025/poster/30844 abstract： Data heterogeneity and backdoor attacks rank among the most significant challenges facing federated learning (FL). For data heterogeneity, personalized federated learning (PFL) enables each client to maintain a private personalized model to cater to client-specific knowledge. Meanwhile, vanilla FL has proven vulnerable to backdoor attacks. However, recent advancements in PFL community have demonstrated a potential immunity against such attacks. This paper explores this intersection further, revealing that existing federated backdoor attacks fail in PFL because backdoors about manually designed triggers struggle to survive in personalized models. To tackle this, we degisn Bad-PFL, which employs features from natural data as our trigger. As long as the model is trained on natural data, it inevitably embeds the backdoor associated with our trigger, ensuring its longevity in personalized models. Moreover, our trigger undergoes mutual reinforcement training with the model, further solidifying the backdoor's durability and enhancing attack effectiveness. The large-scale experiments across three benchmark datasets demonstrate the superior performance of Bad-PFL against various PFL methods, even when equipped with state-of-the-art defense mechanisms.

224. Deep Linear Probe Generators for Weight Space Learning

链接：https://iclr.cc/virtual/2025/poster/29289 abstract： Weight space learning aims to extract information about a neural network, such as its training dataset or generalization error. Recent approaches learn directly from model weights, but this presents many challenges as weights are high-dimensional and include permutation symmetries between neurons. An alternative approach, Probing, represents a model by passing a set of learned inputs (probes) through the model, and training a predictor on top of the corresponding outputs. Although probing is typically not used as a stand alone approach, our preliminary experiment found that a vanilla probing baseline worked surprisingly well. However, we discover that current probe learning strategies are ineffective. We therefore propose Deep Linear Probe Generators (ProbeGen), a simple and effective modification to probing approaches. ProbeGen adds a shared generator module with a deep linear architecture, providing an inductive bias towards structured probes thus reducing overfitting. While simple, ProbeGen performs significantly better than the state-of-the-art and is very efficient, requiring between 30 to 1000 times fewer FLOPs than other top approaches.

225. UNSURE: self-supervised learning with Unknown Noise level and Stein's Unbiased Risk Estimate

链接：https://iclr.cc/virtual/2025/poster/29592 abstract： Recently, many self-supervised learning methods for image reconstruction have been proposed that can learn from noisy data alone, bypassing the need for ground-truth references. Most existing methods cluster around two classes: i) Stein's Unbiased Risk Estimate (SURE) and similar approaches that assume full knowledge of the noise distribution, and ii) Noise2Self and similar cross-validation methods that require very mild knowledge about the noise distribution. The first class of methods tends to be impractical, as the noise level is often unknown in real-world applications, and the second class is often suboptimal compared to supervised learning.In this paper, we provide a theoretical framework that characterizes this expressivity-robustness trade-off and propose a new approach based on SURE, but unlike the standard SURE, does not require knowledge about the noise level. Throughout a series of experiments, we show that the proposed estimator outperforms other existing self-supervised methods on various imaging inverse problems.

226. Preserving Deep Representations in One-Shot Pruning: A Hessian-Free Second-Order Optimization Framework

链接：https://iclr.cc/virtual/2025/poster/28931 abstract： We present SNOWS, a one-shot post-training pruning framework aimed at reducing the cost of vision network inference without retraining. Current leading one-shot pruning methods minimize layer-wise least squares reconstruction error which does not take into account deeper network representations. We propose to optimize a more global reconstruction objective. This objective accounts for nonlinear activations deep in the network to obtain a better proxy for the network loss. This nonlinear objective leads to a more challenging optimization problem---we demonstrate it can be solved efficiently using a specialized second-order optimization framework. A key innovation of our framework is the use of Hessian-free optimization to compute exact Newton descent steps without needing to compute or store the full Hessian matrix. A distinct advantage of SNOWS is that it can be readily applied on top of any sparse mask derived from prior methods, readjusting their weights to preserve deep feature representations. SNOWS obtains state-of-the-art results on various one-shot pruning benchmarks including residual networks and Vision Transformers (ViT/B-16 and ViT/L-16, 86m and 304m parameters respectively). Our open-source implementation is available at https://github.com/mazumder-lab/SNOWS.

227. Wicked Oddities: Selectively Poisoning for Effective Clean-Label Backdoor Attacks

链接：https://iclr.cc/virtual/2025/poster/31195 abstract： Deep neural networks are vulnerable to backdoor attacks, a type of adversarial attack that poisons the training data to manipulate the behavior of models trained on such data. Clean-label backdoor is a more stealthy form of backdoor attacks that can perform the attack without changing the labels of poisoned data.Early works on clean-label attacks added triggers to a random subset of the training set, ignoring the fact that samples contribute unequally to the attack's success. This results in high poisoning rates and low attack success rates.To alleviate the problem, several supervised learning-based sample selection strategies have been proposed.However, these methods assume access to the entire labeled training set and require training, which is expensive and may not always be practical.This work studies a new and more practical (but also more challenging) threat model where the attacker only provides data for the target class (e.g., in face recognition systems) and has no knowledge of the victim model or any other classes in the training set.We study different strategies for selectively poisoning a small set of training samples in the target class to boost the attack success rate in this setting. Our threat model poses a serious threat in training machine learning models with third-party datasets, since the attack can be performed effectively with limited information. Experiments on benchmark datasets illustrate the effectiveness of our strategies in improving clean-label backdoor attacks.

228. Graph Neural Networks Are More Than Filters: Revisiting and Benchmarking from A Spectral Perspective

链接：https://iclr.cc/virtual/2025/poster/28406 abstract： Graph Neural Networks (GNNs) have achieved remarkable success in various graph-based learning tasks. While their performance is often attributed to the powerful neighborhood aggregation mechanism, recent studies suggest that other components such as non-linear layers may also significantly affecting how GNNs process the input graph data in the spectral domain. Such evidence challenges the prevalent opinion that neighborhood aggregation mechanisms dominate the behavioral characteristics of GNNs in the spectral domain. To demystify such a conflict, this paper introduces a comprehensive benchmark to measure and evaluate GNNs' capability in capturing and leveraging the information encoded in different frequency components of the input graph data. Specifically, we first conduct an exploratory study demonstrating that GNNs can flexibly yield outputs with diverse frequency components even when certain frequencies are absent or filtered out from the input graph data. We then formulate a novel research problem of measuring and benchmarking the performance of GNNs from a spectral perspective. To take an initial step towards a comprehensive benchmark, we design an evaluation protocol supported by comprehensive theoretical analysis. Finally, we introduce a comprehensive benchmark on real-world datasets, revealing insights that challenge prevalent opinions from a spectral perspective. We believe that our findings will open new avenues for future advancements in this area. Our implementations can be found at: https://github.com/yushundong/Spectral-benchmark.

229. Scaling Transformers for Low-Bitrate High-Quality Speech Coding

链接：https://iclr.cc/virtual/2025/poster/31006 abstract： The tokenization of audio with neural audio codec models is a vital part of modern AI pipelines for the generation or understanding of speech, alone or in a multimodal context. Traditionally such tokenization models have concentrated on low parameter-count architectures using only components with strong inductive biases. In this work we show that by applying a transformer architecture with large parameter count to this problem, and applying a flexible Finite Scalar Quantization (FSQ) based bottleneck, it is possible to reach state-of-the-art speech quality at extremely low bit-rates of 400 or 700 bits-per-second. The trained models strongly out-perform existing baselines in both objective and subjective tests.

230. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge

链接：https://iclr.cc/virtual/2025/poster/31088 abstract： LLM-as-a-Judge has been widely utilized as an evaluation method in various benchmarks and served as supervised rewards in model training. However, despite their excellence in many domains, potential issues are under-explored, undermining their reliability and the scope of their utility. Therefore, we identify 12 key potential biases and propose a new automated bias quantification framework—CALM—which systematically quantifies and analyzes each type of bias in LLM-as-a-Judge by using automated and principle-guided modification. Our experiments cover multiple popular language models, and the results indicate that while advanced models have achieved commendable overall performance, significant biases persist in certain specific tasks. Empirical results suggest that there remains room for improvement in the reliability of LLM-as-a-Judge. Moreover, we also discuss the explicit and implicit influence of these biases and give some suggestions for the reliable application of LLM-as-a-Judge. Our work highlights the need for stakeholders to address these issues and remind users to exercise caution in LLM-as-a-Judge applications.

231. DataGen: Unified Synthetic Dataset Generation via Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30365 abstract： Large Language Models (LLMs) such as GPT-4 and Llama3 have significantly impacted various fields by enabling high-quality synthetic data generation and reducing dependence on expensive human-generated datasets. Despite this, challenges remain in the areas of generalization, controllability, diversity, and truthfulness within the existing generative frameworks. To address these challenges, this paper presents DataGen, a comprehensive LLM-powered framework designed to produce diverse, accurate, and highly controllable datasets. DataGen is adaptable, supporting all types of text datasets and enhancing the generative process through innovative mechanisms. To augment data diversity, DataGen incorporates an attribute-guided generation module and a group checking feature. For

accuracy, it employs a code-based mathematical assessment for label verification alongside a retrieval-augmented generation technique for factual validation. The framework also allows for user-specified constraints, enabling customization of the data generation process to suit particular requirements. Extensive experiments demonstrate the superior quality of data generated by DataGen, and each module within DataGen plays a critical role in this enhancement. Additionally, DataGen is applied in two practical scenarios: benchmarking LLMs and data augmentation. The results indicate that DataGen effectively supports dynamic and evolving benchmarking and that data augmentation improves LLM capabilities in various domains, including agent-oriented abilities and reasoning skills.

232. GUI-World: A Video Benchmark and Dataset for Multimodal GUI-oriented Understanding

链接：https://iclr.cc/virtual/2025/poster/32086 abstract： Recently, Multimodal Large Language Models (MLLMs) have been used as agents to control keyboard and mouse inputs by directly perceiving the Graphical User Interface (GUI) and generating corresponding commands.However, current agents primarily demonstrate strong understanding capabilities in static environments and are mainly applied to relatively simple domains, such as Web or mobile interfaces.We argue that a robust GUI agent should be capable of perceiving temporal information on the GUI, including dynamic Web content and multi-step tasks.Additionally, it should possess a comprehensive understanding of various GUI scenarios, including desktop software and multi-window interactions.To this end, this paper introduces a new dataset, termed GUI-World, which features meticulously crafted Human-MLLM annotations, extensively covering six GUI scenarios and eight types of GUI-oriented questions in three formats.We evaluate the capabilities of current state-of-the-art MLLMs, including Image LLMs and Video LLMs, in understanding various types of GUI content, especially dynamic and sequential content. Our findings reveal that current models struggle with dynamic GUI content without manually annotated keyframes or operation history. On the other hand, Video LLMs fall short in all GUI-oriented tasks given the sparse GUI video dataset. Therefore, we take the initial step of leveraging a fine-tuned Video LLM, GUI-Vid, as a GUI-oriented assistant, demonstrating an improved understanding of various GUI tasks. However, due to the limitations in the performance of base LLMs, we conclude that using video LLMs as GUI agents remains a significant challenge. We believe our work provides valuable insights for future research in dynamic GUI content understanding. All the dataset and code are publicly available at: https://gui-world.github.io.

233. CR-CTC: Consistency regularization on CTC for improved speech recognition

链接：https://iclr.cc/virtual/2025/poster/30523 abstract： Connectionist Temporal Classification (CTC) is a widely used method for automatic speech recognition (ASR), renowned for its simplicity and computational efficiency. However, it often falls short in recognition performance. In this work, we propose the Consistency-Regularized CTC (CR-CTC), which enforces consistency between two CTC distributions obtained from different augmented views of the input speech mel-spectrogram. We provide in-depth insights into its essential behaviors from three perspectives: 1) it conducts self-distillation between random pairs of sub-models that process different augmented views; 2) it learns contextual representation through masked prediction for positions within time-masked regions, especially when we increase the amount of time masking; 3) it suppresses the extremely peaky CTC distributions, thereby reducing overfitting and improving the generalization ability. Extensive experiments on LibriSpeech, Aishell-1, and GigaSpeech datasets demonstrate the effectiveness of our CR-CTC. It significantly improves the CTC performance, achieving state-of-the-art results comparable to those attained by transducer or systems combining CTC and attention-based encoder-decoder (CTC/AED). We release our code at \url{https://github.com/k2-fsa/icefall}.

234. DEEM: Diffusion models serve as the eyes of large language models for image perception

链接：https://iclr.cc/virtual/2025/poster/28231 abstract： The development of large language models (LLMs) has significantly advanced the emergence of large multimodal models (LMMs). While LMMs have achieved tremendous success by promoting the synergy between multimodal comprehension and creation, they often face challenges when confronted with out-of-distribution data, such as which can hardly distinguish orientation, quantity, color, structure, etc. This is primarily due to their reliance on image encoders trained to encode images into task-relevant features, which may lead them to disregard irrelevant details. Delving into the modeling capabilities of diffusion models for images naturally prompts the question: Can diffusion models serve as the eyes of large language models for image perception? In this paper, we propose DEEM, a simple but effective approach that utilizes the generative feedback of diffusion models to align the semantic distributions of the image encoder. This addresses the drawbacks of previous methods that solely relied on image encoders like CLIP-ViT, thereby enhancing the model's resilience against out-of-distribution samples and reducing visual hallucinations. Importantly, this is achieved without requiring additional training modules and with fewer training parameters. We extensively evaluated DEEM on both our newly constructed RobustVQA benchmark and other well-known benchmarks, POPE and MMVP, for visual hallucination and perception. In particular, DEEM improves LMM's visual perception performance to a large extent (e.g., 4\% ↑ on RobustVQA, 6.5\% ↑ on MMVP and 12.8 \% ↑ on POPE). Compared to the state-of-the-art interleaved content generation models, DEEM exhibits enhanced robustness and a superior capacity to alleviate model hallucinations while utilizing fewer trainable parameters, less pre-training data (10\%), and a smaller base model size. Extensive experiments demonstrate that DEEM enhances the performance of LMMs on various downstream tasks without inferior performance in the long term, including visual question answering, image captioning, and text-conditioned image synthesis.

235. TS-LIF: A Temporal Segment Spiking Neuron Network for Time Series Forecasting

链接：https://iclr.cc/virtual/2025/poster/28210 abstract： Spiking Neural Networks (SNNs) offer a promising, biologically inspired approach for processing spatiotemporal data, particularly for time series forecasting.However, conventional neuron models like the Leaky Integrate-and-Fire (LIF) struggle to capture long-term dependencies and effectively process multi-scale temporal dynamics.To overcome these limitations, we introduce the Temporal Segment Leaky Integrate-and-Fire (TS-LIF) model, featuring a novel dual-compartment architecture.The dendritic and somatic compartments specialize in capturing distinct frequency components, providing functional heterogeneity that enhances the neuron's ability to process both low- and high-frequency information.Furthermore, the newly introduced direct somatic current injection reduces information loss during intra-neuronal transmission, while dendritic spike generation improves multi-scale information extraction.We provide a theoretical stability analysis of the TS-LIF model and explain how each compartment contributes to distinct frequency response characteristics.Experimental results show that TS-LIF outperforms traditional SNNs in time series forecasting, demonstrating better accuracy and robustness, even with missing data.TS-LIF advances the application of SNNs in time-series forecasting, providing a biologically inspired approach that captures complex temporal dynamics and offers potential for practical implementation in diverse forecasting scenarios.

236. FLOPS: Forward Learning with OPtimal Sampling

链接：https://iclr.cc/virtual/2025/poster/27694 abstract： Given the limitations of backpropagation, perturbation-based gradient computation methods have recently gained focus for learning with only forward passes, also referred to as queries. Conventional forward learning consumes enormous queries on each data point for accurate gradient estimation through Monte Carlo sampling, which hinders the scalability of those algorithms. However, not all data points deserve equal queries for gradient estimation. In this paper, we study the problem of improving the forward learning efficiency from a novel perspective: how to reduce the gradient estimation variance with minimum cost? For this, we allocate the optimal number of queries within a set budget during training to balance estimation accuracy and computational efficiency. Specifically, with a simplified proxy objective and a reparameterization technique, we derive a novel plug-and-play query allocator with minimal parameters. Theoretical results are carried out to verify its optimality. We conduct extensive experiments for fine-tuning Vision Transformers on various datasets and further deploy the allocator to two black-box applications: prompt tuning and multimodal alignment for foundation models. All findings demonstrate that our proposed allocator significantly enhances the scalability of forward-learning algorithms, paving the way for real-world applications. The implementation is available at https://github.com/RTkenny/FLOPS-Forward-Learning-with-OPtimal-Sampling.

237. AdvPaint: Protecting Images from Inpainting Manipulation via Adversarial Attention Disruption

链接：https://iclr.cc/virtual/2025/poster/28482 abstract： The outstanding capability of diffusion models in generating high-quality images poses significant threats when misused by adversaries. In particular, we assume malicious adversaries exploiting diffusion models for inpainting tasks, such as replacing a specific region with a celebrity. While existing methods for protecting images from manipulation in diffusion-based generative models have primarily focused on image-to-image and text-to-image tasks, the challenge of preventing unauthorized inpainting has been rarely addressed, often resulting in suboptimal protection performance. To mitigate inpainting abuses, we propose ADVPAINT, a novel defensive framework that generates adversarial perturbations that effectively disrupt the adversary’s inpainting tasks. ADVPAINT targets the self- and cross-attention blocks in a target diffusion inpainting model to distract semantic understanding and prompt interactions during image generation. ADVPAINT also employs a two-stage perturbation strategy, dividing the perturbation region based on an enlarged bounding box around the object, enhancing robustness across diverse masks of varying shapes and sizes. Our experimental results demonstrate that ADVPAINT’s perturbations are highly effective in disrupting the adversary’s inpainting tasks, outperforming existing methods; ADVPAINT attains over a 100-point increase in FID and substantial decreases in precision.

238. An Illustrated Guide to Automatic Sparse Differentiation

链接：https://iclr.cc/virtual/2025/poster/31324 abstract： In numerous applications of machine learning, Hessians and Jacobians exhibit sparsity, a property that can be leveraged to vastly accelerate their computation. While the usage of automatic differentiation in machine learning is ubiquitous, automatic sparse differentiation (ASD) remains largely unknown. This post introduces ASD, explaining its key components and their roles in the computation of both sparse Jacobians and Hessians. We conclude with a practical demonstration showcasing the performance benefits of ASD._First-order optimization is ubiquitous in machine learning (ML) but second-order optimization is much less common. The intuitive reason is that high-dimensional vectors (gradients) are cheap, whereas high-dimensional matrices (Hessians) are expensive. Luckily, in numerous applications of ML to science or engineering, Hessians and Jacobians exhibit sparsity: most of their coefficients are known to be zero. Leveraging this sparsity can vastly accelerate automatic differentiation (AD) for Hessians and Jacobians, while decreasing its memory requirements. Yet, while traditional AD is available in many high-level programming languages like Python and Julia, automatic sparse differentiation (ASD) is not as widely used. One reason is that the underlying theory was developed outside of the ML research ecosystem, by people more familiar with low-level programming languages.With this blog post, we aim to shed light on the inner workings of ASD, bridging the gap between the ML and AD communities by presenting well established techniques from the latter field. We start out with a short introduction to traditional AD, covering the computation of Jacobians in both forward

and reverse mode. We then dive into the two primary components of ASD: sparsity pattern detection and matrix coloring. Having described the computation of sparse Jacobians, we move on to sparse Hessians.We conclude with a practical demonstration of ASD, providing performance benchmarks and guidance on when to use ASD over AD.

239. RGB-Event ISP: The Dataset and Benchmark

链接：https://iclr.cc/virtual/2025/poster/30541 abstract： Event-guided imaging has received significant attention due to its potential to revolutionize instant imaging systems. However, the prior methods primarily focus on enhancing RGB images in a post-processing manner, neglecting the challenges of image signal processor (ISP) dealing with event sensor and the benefits events provide for reforming the ISP process. To achieve this, we conduct the first research on event-guided ISP. First, we present a new event-RAW paired dataset, collected with a novel but still confidential sensor that records pixel-level aligned events and RAW images. This dataset includes 3373 RAW images with 2248×3264 resolution and their corresponding events, spanning 24 scenes with 3 exposure modes and 3 lenses. Second, we propose a convential ISP pipeline to generate good RGB frames as reference. This convential ISP pipleline performs basic ISP operations, e.g., demosaicing, white balancing, denoising and color space transforming, with a ColorChecker as reference. Third, we classify the existing learnable ISP methods into 3 classes, and select multiple methods to train and evaluate on our new dataset. Lastly, since there is no prior work for reference, we propose a simple event-guided ISP method and test it on our dataset. We further put forward key technical challenges and future directions in RGB-Event ISP. In summary, to the best of our knowledge, this is the very first research focusing on event-guided ISP, and we hope it will inspire the community.

240. Animate-X: Universal Character Image Animation with Enhanced Motion Representation

链接：https://iclr.cc/virtual/2025/poster/31209 abstract： Character image animation, which generates high-quality videos from a reference image and target pose sequence, has seen significant progress in recent years. However, most existing methods only apply to human figures, which usually do not generalize well on anthropomorphic characters commonly used in industries like gaming and entertainment. Our in-depth analysis suggests to attribute this limitation to their insufficient modeling of motion, which is unable to comprehend the movement pattern of the driving video, thus imposing a pose sequence rigidly onto the target character. To this end, this paper proposes $\texttt{Animate-X}$, a universal animation framework based on LDM for various character types (collectively named $\texttt{X}$), including anthropomorphic characters. To enhance motion representation, we introduce the Pose Indicator, which captures comprehensive motion pattern from the driving video through both implicit and explicit manner. The former leverages CLIP visual features of a driving video to extract its gist of motion, like the overall movement pattern and temporal relations among motions, while the latter strengthens the generalization of LDM by simulating possible inputs in advance that may arise during inference. Moreover, we introduce a new Animated Anthropomorphic Benchmark ($\texttt{$A^2$Bench}$) to evaluate the performance of $\texttt{Animate-X}$ on universal and widely applicable animation images. Extensive experiments demonstrate the superiority and effectiveness of $\texttt{Animate-X}$ compared to state-of-the-art methods.

241. SimPER: A Minimalist Approach to Preference Alignment without Hyperparameters

链接：https://iclr.cc/virtual/2025/poster/28626 abstract： Existing preference optimization objectives for language model alignment require additional hyperparameters that must be extensively tuned to achieve optimal performance, increasing both the complexity and time required for fine-tuning large language models. In this paper, we propose a simple yet effective hyperparameter-free preference optimization algorithm for alignment. We observe that promising performance can be achieved simply by optimizing inverse perplexity, which is calculated as the inverse of the exponentiated average log-likelihood of the chosen and rejected responses in the preference dataset. The resulting simple learning objective, SimPER, is easy to implement and eliminates the need for expensive hyperparameter tuning and a reference model, making it both computationally and memory efficient. Extensive experiments on widely used real-world benchmarks, including MT-Bench, AlpacaEval 2, and 10 key benchmarks of the Open LLM Leaderboard with 5 base models, demonstrate that SimPER consistently and significantly outperforms existing approaches—even without any hyperparameters or a reference model. For example, despite its simplicity, SimPER outperforms state-of-the-art methods by up to 5.7 points on AlpacaEval 2 and achieves the highest average ranking across 10 benchmarks on the Open LLM Leaderboard. The source code for SimPER is publicly available at: https://github.com/tengxiao1/SimPER.

242. Dataset Distillation via Knowledge Distillation: Towards Efficient Self-Supervised Pre-training of Deep Networks

链接：https://iclr.cc/virtual/2025/poster/29070 abstract： Dataset distillation (DD) generates small synthetic datasets that can efficiently train deep networks with a limited amount of memory and compute. Despite the success of DD methods for supervised learning, DD for self-supervised pre-training of deep models has remained unaddressed. Pre-training on unlabeled data is crucial for efficiently generalizing to downstream tasks with limited labeled data. In this work, we propose the first effective DD method for SSL pre-training. First, we show, theoretically and empirically, that naiive application of supervised DD methods to SSL fails, due to the high variance of the SSL gradient. Then, we address this issue by relying on insights from

knowledge distillation (KD) literature. Specifically, we train a small student model to match the representations of a larger teacher model trained with SSL. Then, we generate a small synthetic dataset by matching the training trajectories of the student models. As the KD objective has considerably lower variance than SSL, our approach can generate synthetic datasets that can successfully pre-train high-quality encoders. Through extensive experiments, we show that our distilled sets lead to up to 13% higher accuracy than prior work, on a variety of downstream tasks, in the presence of limited labeled data. Code at https://github.com/BigML-CS-UCLA/MKDT.

243. OmniKV: Dynamic Context Selection for Efficient Long-Context LLMs

链接：https://iclr.cc/virtual/2025/poster/27961 abstract： During the inference phase of Large Language Models (LLMs) with long context, a substantial portion of GPU memory is allocated to the KV cache, with memory usage increasing as the sequence length grows. To mitigate the GPU memory footprint associate with KV cache, some previous studies have discarded less important tokens based on the sparsity identified in attention scores in long context scenarios. However, we argue that attention scores cannot indicate the future importance of tokens in subsequent generation iterations, because attention scores are calculated based on current hidden states. Therefore, we propose OmniKV, a token-dropping-free and training-free inference method, which achieves a 1.68x speedup without any loss in performance. It is well-suited for offloading, significantly reducing KV cache memory usage by up to 75% with it. The core innovative insight of OmniKV is: Within a single generation iteration, there is a high degree of similarity in the important tokens identified across consecutive layers. Extensive experiments demonstrate that OmniKV achieves state-of-the-art performance across multiple benchmarks, with particularly advantages in chain-of-thoughts scenarios. OmniKV extends the maximum context length supported by a single A100 for Llama-3-8B from 128K to 450K. Our code is available at https://github.com/antgroup/OmniKV.git.

244. Inverse Constitutional AI: Compressing Preferences into Principles

链接：https://iclr.cc/virtual/2025/poster/30711 abstract： Feedback data is widely used for fine-tuning and evaluating state-of-the-art AI models. Pairwise text preferences, where human or AI annotators select the “better” of two options, are particularly common. Such preferences are used to train (reward) models or to rank models with aggregate statistics. For many applications it is desirable to understand annotator preferences in addition to modelling them – not least because extensive prior work has shown various unintended biases in preference datasets. Yet, preference datasets remain challenging to interpret. Neither black-box reward models nor statistics can answer why one text is preferred over another. Manual interpretation of the numerous (long) response pairs is usually equally infeasible. In this paper, we introduce the Inverse Constitutional AI (ICAI) problem, formulating the interpretation of pairwise text preference data as a compression task. In constitutional AI, a set of principles (a constitution) is used to provide feedback and fine-tune AI models. ICAI inverts this process: given a feedback dataset, we aim to extract a constitution that best enables a large language model (LLM) to reconstruct the original annotations. We propose a corresponding ICAI algorithm and validate its generated constitutions quantitatively based on annotation reconstruction accuracy on several datasets: (a) synthetic feedback data with known principles; (b) AlpacaEval cross-annotated human feedback data; (c) crowdsourced Chatbot Arena data; and (d) PRISM data from diverse demographic groups. As an example application, we further demonstrate the detection of biases in human feedback data. As a short and interpretable representation of the original dataset, generated constitutions have many potential use cases: they may help identify undesirable annotator biases, better understand model performance, scale feedback to unseen data, or assist with adapting AI models to individual user or group preferences. We release the source code for our algorithm and experiments at https://github.com/rdnfn/icai.

245. Neural Functions for Learning Periodic Signal

链接：https://iclr.cc/virtual/2025/poster/30294 abstract： As function approximators, deep neural networks have served as an effective tool to represent various signal types. Recent approaches utilize multi-layer perceptrons (MLPs) to learn a nonlinear mapping from a coordinate to its corresponding signal, facilitating the learning of continuous neural representations from discrete data points. Despite notable successes in learning diverse signal types, coordinate-based MLPs often face issues of overfitting and limited generalizability beyond the training region, resulting in subpar extrapolation performance. This study addresses scenarios where the underlying true signals exhibit periodic properties, either spatially or temporally. We propose a novel network architecture, which extracts periodic patterns from measurements and leverages this information to represent the signal, thereby enhancing generalization and improving extrapolation performance. We demonstrate the efficacy of the proposed method through comprehensive experiments, including the learning of the periodic solutions for differential equations, and time series imputation (interpolation) and forecasting (extrapolation) on real-world datasets.

246. MACPO: Weak-to-Strong Alignment via Multi-Agent Contrastive Preference Optimization

链接：https://iclr.cc/virtual/2025/poster/27806 abstract： As large language models (LLMs) are rapidly advancing and achieving near-human capabilities on specific tasks, aligning them with human values is becoming more urgent. In scenarios where LLMs outperform humans, we face a weak-to-strong alignment problem where we need to effectively align strong student LLMs through weak supervision generated by weak teachers. Existing alignment methods mainly focus on strong-to-weak alignment and self-alignment settings, and it is impractical to adapt them to the much harder weak-to-strong alignment setting. To fill this gap, we propose a multi-agent contrastive preference optimization (MACPO) framework. MACPO facilitates weak teachers and strong students to learn from each other by iteratively reinforcing unfamiliar positive behaviors while penalizing familiar negative ones. To get this, we devise a mutual positive behavior augmentation strategy to encourage weak teachers and strong students

to learn from each other's positive behavior and further provide higher quality positive behavior for the next iteration. Additionally, we propose a hard negative behavior construction strategy to induce weak teachers and strong students to generate familiar negative behavior by fine-tuning on negative behavioral data. Experimental results on the HH-RLHF and PKU-SafeRLHF datasets, evaluated using both automatic metrics and human judgments, demonstrate that MACPO simultaneously improves the alignment performance of strong students and weak teachers. Moreover, as the number of weak teachers increases, MACPO achieves better weak-to-strong alignment performance through more iteration optimization rounds.

247. Infilling Score: A Pretraining Data Detection Algorithm for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30693 abstract： In pretraining data detection, the goal is to detect whether a given sentence is in the dataset used for training a Large Language Model LLM). Recent methods (such as Min-K % and Min-K%++) reveal that most training corpora are likely contaminated with both sensitive content and evaluation benchmarks, leading to inflated test set performance. These methods sometimes fail to detect samples from the pretraining data, primarily because they depend on statistics composed of causal token likelihoods. We introduce Infilling Score, a new test-statistic based on non-causal token likelihoods. Infilling Score can be computed for autoregressive models without re-training using Bayes rule. A naive application of Bayes rule scales linearly with the vocabulary size. However, we propose a ratio test-statistic whose computation is invariant to vocabulary size. Empirically, our method achieves a significant accuracy gain over state-of-the-art methods including Min-K%, and Min-K%++ on the WikiMIA benchmark across seven models with different parameter sizes. Further, we achieve higher AUC compared to reference-free methods on the challenging MIMIR benchmark. Finally, we create a benchmark dataset consisting of recent data sources published after the release of Llama-3; this benchmark provides a statistical baseline to indicate potential corpora used for Llama-3 training.

248. Measuring And Improving Persuasiveness Of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29866 abstract： Large Language Models (LLMs) are increasingly being used in workflows involving generating content to be consumed by humans (e.g., marketing) and also in directly interacting with humans (e.g., through chatbots). The development of such systems that are capable of generating verifiably persuasive messages presents both opportunities and challenges for society. On the one hand, such systems could positively impact domains like advertising and social good, such as addressing drug addiction, and on the other, they could be misused for spreading misinformation and shaping political opinions. To channel LLMs' impact on society, we need to develop systems to measure and benchmark their persuasiveness. With this motivation, we introduce PersuasionBench and PersuasionArena, the first large-scale benchmark and arena containing a battery of tasks to automatically measure the simulative and generative persuasion abilities of large language models. We introduce transsuasion (trans = carrying across, suasion = the act of persuading), a novel task of transforming non-persuasive language into persuasive content while preserving other factors determining persuasiveness (sender, receiver, time, and channel). Our findings indicate that the simulative persuasion capabilities of LLMs are barely above random; however, their generative persuasion capabilities are much better. For instance, GPT-4o loses only 36% of the time when playing against the best human persuader. Further, we find that LLMs' persuasiveness correlates positively with model size, but smaller models can also be made to have a higher persuasiveness than much larger models. Notably, targeted training using synthetic and natural datasets significantly enhances smaller models' persuasive capabilities, challenging scale-dependent assumptions. Our findings carry key implications for both model developers and policymakers. For instance, while the EU AI Act and California's SB-1047 aim to regulate AI models based on the number of floating point operations, we demonstrate that simple metrics like this alone fail to capture the full scope of AI's societal impact. We invite the community to explore and contribute to PersuasionArena and PersuasionBench, available at behavior-in-the-wild.github.io/measure-persuasion, to advance our understanding of AI-driven persuasion and its societal implications.

249. Teaching Human Behavior Improves Content Understanding Abilities Of VLMs

链接：https://iclr.cc/virtual/2025/poster/28866 abstract： Communication is defined as "Who says what to whom with what effect." A message from a communicator generates downstream receiver effects, also known as behavior. Receiver behavior, being a downstream effect of the message, carries rich signals about it. Even after carrying signals about the message, the behavior signal is often ignored while training vision-language models. We show that training VLMs on receiver behavior can actually help improve their content-understanding abilities. We demonstrate that training VLMs to predict receiver behaviors, such as likes, comments, and replay graphs, which are available at scale, enhances the VLM's performance across a broad range of downstream content understanding tasks. We show this performance increase over 6 types of behavior, 46 different tasks covering image, video, text, and audio over 26 benchmark datasets across both zero-shot and fine-tuning settings, outperforming many supervised baselines on diverse tasks ranging from emotion recognition to captioning by up to 150%. We note that since receiver behavior, such as likes, comments, and replay graphs, is collected by default on the internet and does not need any human annotations to be useful, the performance improvement we get after training on this data is essentially free lunch. We also release BLIFT, our Behaviour-LLaVA IFT dataset comprising 730k images and videos with their receiver behavior collected from multiple platforms on which we train our models to achieve this. The dataset and code are available at behavior-in-the-wild.github.io/behavior-llava.

250. Diff-Prompt: Diffusion-driven Prompt Generator with Mask Supervision

链接：https://iclr.cc/virtual/2025/poster/29982 abstract： Prompt learning has demonstrated promising results in fine-tuning pre-trained multimodal models. However, the performance improvement is limited when applied to more complex and fine-grained tasks. The reason is that most existing methods directly optimize the parameters involved in the prompt generation process through loss backpropagation, which constrains the richness and specificity of the prompt representations. In this paper, we propose Diffusion-Driven Prompt Generator (Diff-Prompt), aiming to use the diffusion model to generate rich and fine-grained prompt information for complex downstream tasks. Specifically, our approach consists of three stages. In the first stage, we train a Mask-VAE to compress the masks into latent space. In the second stage, we leverage an improved Diffusion Transformer (DiT) to train a prompt generator in the latent space, using the masks for supervision. In the third stage, we align the denoising process of the prompt generator with the pre-trained model in the semantic space, and use the generated prompts to fine-tune the model. We conduct experiments on a complex pixel-level downstream task, referring expression comprehension, and compare our method with various parameter-efficient fine-tuning approaches. Diff-Prompt achieves a maximum improvement of 8.87 in R@1 and 14.05 in R@5 compared to the foundation model and also outperforms other state-of-the-art methods across multiple metrics. The experimental results validate the effectiveness of our approach and highlight the potential of using generative models for prompt generation. Code is available at https://github.com/Kelvin-ywc/diff-prompt.

251. Competing Large Language Models in Multi-Agent Gaming Environments

链接：https://iclr.cc/virtual/2025/poster/30468 abstract： Decision-making is a complex process requiring diverse abilities, making it an excellent framework for evaluating Large Language Models (LLMs). Researchers have examined LLMs' decision-making through the lens of Game Theory. However, existing evaluation mainly focus on two-player scenarios where an LLM competes against another. Additionally, previous benchmarks suffer from test set leakage due to their static design. We introduce GAMA(γ)-Bench, a new framework for evaluating LLMs' Gaming Ability in Multi-Agent environments. It includes eight classical game theory scenarios and a dynamic scoring scheme specially designed to quantitatively assess LLMs' performance. γ-Bench allows flexible game settings and adapts the scoring system to different game parameters, enabling comprehensive evaluation of robustness, generalizability, and strategies for improvement. Our results indicate that GPT-3.5 demonstrates strong robustness but limited generalizability, which can be enhanced using methods like Chain-of-Thought. We also evaluate 13 LLMs from 6 model families, including GPT-3.5, GPT-4, Gemini, LLaMA-3.1, Mixtral, and Qwen-2. Gemini-1.5-Pro outperforms others, scoring of 69.8 out of 100, followed by LLaMA-3.1-70B (65.9) and Mixtral-8x22B (62.4). Our code and experimental results are publicly available at https://github.com/CUHK-ARISE/GAMABench.

252. Zeroth-Order Fine-Tuning of LLMs with Transferable Static Sparsity

链接：https://iclr.cc/virtual/2025/poster/28438 abstract： Zeroth-order optimization (ZO) is a memory-efficient strategy for fine-tuning Large Language Models using only forward passes. However, applying ZO fine-tuning in memory-constrained settings such as mobile phones and laptops remains challenging since these settings often involve weight quantization, while ZO requires full-precision perturbation and update. In this study, we address this limitation by combining static sparse ZO fine-tuning with quantization. Our approach transfers a small, static subset (0.1%) of "sensitive" parameters from pre-training to downstream tasks, focusing fine-tuning on this sparse set of parameters. The remaining untuned parameters are quantized, reducing memory demands. Our proposed workflow enables efficient ZO fine-tuning of an Llama2-7B model on a GPU device with less than 8GB of memory while outperforming full model ZO fine-tuning performance and in-context learning.

253. From Commands to Prompts: LLM-based Semantic File System for AIOS

链接：https://iclr.cc/virtual/2025/poster/31152 abstract： Large language models (LLMs) have demonstrated significant potential in the development of intelligent LLM-based agents. However, when users use these agent applications to perform file operations, their interaction with the file system still remains the traditional paradigm: reliant on manual navigation through precise commands. This paradigm poses a bottleneck to the usability of these systems as users are required to navigate complex folder hierarchies and remember cryptic file names. To address this limitation, we propose an LLM-based Semantic File System (LSFS) for prompt-driven file management in LLM Agent Operating System (AIOS). Unlike conventional approaches, LSFS incorporates LLMs to enable users or agents to interact with files through natural language prompts, facilitatingsemantic file management. At the macro-level, we develop a comprehensive API set to achieve semantic file management functionalities, such as semantic file retrieval, file update summarization, and semantic file rollback). At the micro-level, we store files by constructing semantic indexes for them, design and implement syscalls of different semantic operations, e.g., CRUD (create, read, update, delete),group by, join. Our experiments show that LSFS can achieve at least 15% retrieval accuracy improvement with 2.1× higher retrieval speed in the semantic file retrieval task compared with the traditional file system. In the traditional keyword-based file retrieval task (i.e., retrieving by string-matching), LSFS also performs stably well, i.e., over 89% F1-score with improved usability, especially when the keyword conditions become more complex. Additionally, LSFS supports more advanced file management operations, i.e., semantic file rollback and file sharing and achieves 100% success rates in these tasks, further suggesting the capability of LSFS . The code is available at https://github.com/agiresearch/AIOS-LSFS.

254. AutoCGP: Closed-Loop Concept-Guided Policies from Unlabeled

Demonstrations

链接：https://iclr.cc/virtual/2025/poster/30674 abstract： Training embodied agents to perform complex robotic tasks presents significant challenges due to the entangled factors of task compositionality, environmental diversity, and dynamic changes. In this work, we introduce a novel imitation learning framework to train closed-loop concept-guided policies that enhance long-horizon task performance by leveraging discovered manipulation concepts. Unlike methods that rely on predefined skills and human-annotated labels, our approach allows agents to autonomously abstract manipulation concepts from their proprioceptive states, thereby alleviating misalignment due to ambiguities in human semantics and environmental complexity. Our framework comprises two primary components: an Automatic Concept Discovery module that identifies meaningful and consistent manipulation concepts, and a Concept-Guided Policy Learning module that effectively utilizes these manipulation concepts for adaptive task execution, including a Concept Selection Transformer for concept-based guidance and a Concept-Guided Policy for action prediction with the selected concepts. Experiments demonstrate that our approach significantly outperforms baseline methods across a range of tasks and environments, while showcasing emergent consistency in motion patterns associated with the discovered manipulation concepts. Codes are available at: https://github.com/PeiZhou26/AutoCGP.

255. Robust System Identification: Finite-sample Guarantees and Connection to Regularization

链接：https://iclr.cc/virtual/2025/poster/29222 abstract： We consider the problem of learning nonlinear dynamical systems from a single sample trajectory. While the least squares estimate (LSE) is commonly used for this task, it suffers from poor identification errors when the sample size is small or the model fails to capture the system's true dynamics. To overcome these limitations, we propose a robust LSE framework, which incorporates robust optimization techniques, and prove that it is equivalent to regularizing LSE using general Schatten p-norms. We provide non-asymptotic performance guarantees for linear systems, achieving an error rate of $\widetilde{\mathcal{O}}(1/\sqrt{T})$, and show that it avoids the curse of dimensionality, unlike state-of-the-art Wasserstein robust optimization models. Empirical results demonstrate substantial improvements in real-world system identification and online control tasks, outperforming existing methods.

256. Accelerating 3D Molecule Generation via Jointly Geometric Optimal Transport

链接：https://iclr.cc/virtual/2025/poster/29423 abstract： This paper proposes a new 3D molecule generation framework, called GOAT, for fast and effective 3D molecule generation based on the flow-matching optimal transport objective. Specifically, we formulate a geometric transport formula for measuring the cost of mapping multi-modal features (e.g., continuous atom coordinates and categorical atom types) between a base distribution and a target data distribution. Our formula is solved within a joint, equivariant, and smooth representation space. This is achieved by transforming the multi-modal features into a continuous latent space with equivariant networks. In addition, we find that identifying optimal distributional coupling is necessary for fast and effective transport between any two distributions. We further propose a mechanism for estimating and purifying optimal coupling to train the flow model with optimal transport. By doing so, GOAT can turn arbitrary distribution couplings into new deterministic couplings, leading to an estimated optimal transport plan for fast 3D molecule generation. The purification filters out the subpar molecules to ensure the ultimate generation quality. We theoretically and empirically prove that the proposed optimal coupling estimation and purification yield transport plan with non-increasing cost. Finally, extensive experiments show that GOAT enjoys the efficiency of solving geometric optimal transport, leading to a double speedup compared to the sub-optimal method while achieving the best generation quality regarding validity, uniqueness, and novelty.

257. Spurious Forgetting in Continual Learning of Language Models

链接：https://iclr.cc/virtual/2025/poster/29593 abstract： Recent advancements in large language models (LLMs) reveal a perplexing phenomenon in continual learning: despite extensive training, models experience significant performance declines, raising questions about task alignment and underlying knowledge retention. This study first explores the concept of "spurious forgetting", proposing that such performance drops often reflect a decline in task alignment rather than true knowledge loss. Through controlled experiments with a synthesized dataset, we investigate the dynamics of model performance during the initial training phases of new tasks, discovering that early optimization steps can disrupt previously established task alignments. Our theoretical analysis connects these shifts to orthogonal updates in model weights, providing a robust framework for understanding this behavior. Ultimately, we introduce a Freezing strategy that fix the bottom layers of the model, leading to substantial improvements in four continual learning scenarios. Our findings underscore the critical distinction between task alignment and knowledge retention, paving the way for more effective strategies in continual learning.

258. Training Large Language Models for Retrieval-Augmented Question Answering through Backtracking Correction

链接：https://iclr.cc/virtual/2025/poster/30171 abstract： Despite recent progress in Retrieval-Augmented Generation (RAG) achieved by large language models (LLMs), retrievers often recall uncorrelated documents, regarded as "noise" during subsequent text generation. To address this, some methods train LLMs to distinguish between relevant and irrelevant documents using labeled data, enabling them to select the most likely relevant ones as context. However, they remain sensitive

to noise, as LLMs can easily make mistakes when the selected document is noisy. Some approaches increase the number of referenced documents and train LLMs to perform stepwise reasoning when presented with multiple documents. Unfortunately, these methods rely on extensive and diverse annotations to ensure generalization, which is both challenging and costly. In this paper, we propose Backtracking Correction to address these limitations. Specifically, we reformulate stepwise RAG into a multi-step decision-making process. Starting from the final step, we optimize the model through error sampling and self-correction, and then backtrack to the previous state iteratively. In this way, the model's learning scheme follows an easy-to-hard progression: as the target state moves forward, the context space decreases while the decision space increases. Experimental results demonstrate that Backtracking Correction enhances LLMs' ability to make complex multi-step assessments, improving the robustness of RAG in dealing with noisy documents.

259. Transition Path Sampling with Improved Off-Policy Training of Diffusion Path Samplers

链接：https://iclr.cc/virtual/2025/poster/29361 abstract： Understanding transition pathways between two meta-stable states of a molecular system is crucial to advance drug discovery and material design. However, unbiased molecular dynamics (MD) simulations are computationally infeasible because of the high energy barriers that separate these states. Although recent machine learning techniques are proposed to sample rare events, they are often limited to simple systems and rely on collective variables (CVs) derived from costly domain expertise. In this paper, we introduce a novel approach that trains diffusion path samplers (DPS) to address the transition path sampling (TPS) problem without requiring CVs. We reformulate the problem as an amortized sampling from the transition path distribution by minimizing the log-variance divergence between the path distribution induced by DPS and the transition path distribution. Based on the log-variance divergence, we propose learnable control variates to reduce the variance of gradient estimators and the off-policy training objective with replay buffers and simulated annealing techniques to improve sample efficiency and diversity. We also propose a scale-based equivariant parameterization of the bias forces to ensure scalability for large systems. We extensively evaluate our approach, termed TPS-DPS, on a synthetic system, small peptide, and challenging fast-folding proteins, demonstrating that it produces more realistic and diverse transition pathways than existing baselines. We also provide links to project page and code.

260. Tight Lower Bounds under Asymmetric High-Order Hölder Smoothness and Uniform Convexity

链接：https://iclr.cc/virtual/2025/poster/28882 abstract： In this paper, we provide tight lower bounds for the oracle complexity of minimizing high-order Hölder smooth and uniformly convex functions. Specifically, for a function whose p^{th}-order derivatives are Hölder continuous with degree ν and parameter H, and that is uniformly convex with degree q and parameter σ, we focus on two asymmetric cases: (1) $q > p + \nu$, and (2) $q < p+\nu$. Given up to p^{th}-order oracle access, we establish worst-case oracle complexities of $\Omega\left(\left(\frac{H}{\sigma}\right)^\frac{2}{3(p+\nu)-2}\left(\frac{\sigma} {\epsilon}\right)^\frac{2(q-p-\nu)}{q(3(p+\nu)-2)}\right)$ in the first case with an ℓ_∞-ball-truncated-Gaussian smoothed hard function and $\Omega\left(\left(\frac{H}{\sigma}\right)^\frac{2}{3(p+\nu)-2}+ \log\log\left(\left(\frac{\sigma^{p+\nu}} {H^q}\right)^\frac{1}{p+\nu-q}\frac{1}{\epsilon}\right)\right)$ in the second case, for reaching an ϵ-approximate solution in terms of the optimality gap. Our analysis generalizes previous lower bounds for functions under first- and second-order smoothness as well as those for uniformly convex functions, and furthermore our results match the corresponding upper bounds in this general setting.

261. Generalization Bounds and Model Complexity for Kolmogorov–Arnold Networks

链接：https://iclr.cc/virtual/2025/poster/28263 abstract： Kolmogorov–Arnold Network (KAN) is a network structure recently proposed in Liu et al. (2024) that offers improved interpretability and a more parsimonious design in many science-oriented tasks compared to multi-layer perceptrons. This work provides a rigorous theoretical analysis of KAN by establishing generalization bounds for KAN equipped with activation functions that are either represented by linear combinations of basis functions or lying in a low-rank Reproducing Kernel Hilbert Space (RKHS). In the first case, the generalization bound accommodates various choices of basis functions in forming the activation functions in each layer of KAN and is adapted to different operator norms at each layer. For a particular choice of operator norms, the bound scales with the l_1 norm of the coefficient matrices and the Lipschitz constants for the activation functions, and it has no dependence on combinatorial parameters (e.g., number of nodes) outside of logarithmic factors. Moreover, our result does not require the boundedness assumption on the loss function and, hence, is applicable to a general class of regression-type loss functions. In the low-rank case, the generalization bound scales polynomially with the underlying ranks as well as the Lipschitz constants of the activation functions in each layer. These bounds are empirically investigated for KANs trained with stochastic gradient descent on simulated and real data sets. The numerical results demonstrate the practical relevance of these bounds.

262. Discrete Diffusion Schrödinger Bridge Matching for Graph Transformation

链接：https://iclr.cc/virtual/2025/poster/28054 abstract： Transporting between arbitrary distributions is a fundamental goal in generative modeling.Recently proposed diffusion bridge models provide a potential solution, but they rely on a joint distribution

that is difficult to obtain in practice.Furthermore, formulations based on continuous domains limit their applicability to discrete domains such as graphs.To overcome these limitations, we propose Discrete Diffusion Schrödinger Bridge Matching (DDSBM), a novel framework that utilizes continuous-time Markov chains to solve the SB problem in a high-dimensional discrete state space.Our approach extends Iterative Markovian Fitting to discrete domains, and we have proved its convergence to the SB.Furthermore, we adapt our framework for the graph transformation, and show that our design choice of underlying dynamics characterized by independent modifications of nodes and edges can be interpreted as the entropy-regularized version of optimal transport with a cost function described by the graph edit distance.To demonstrate the effectiveness of our framework, we have applied DDSBM to molecular optimization in the field of chemistry.Experimental results demonstrate that DDSBM effectively optimizes molecules' property-of-interest with minimal graph transformation, successfully retaining other features. Source code is available here.

263. CyberHost: A One-stage Diffusion Framework for Audio-driven Talking Body Generation

链接：https://iclr.cc/virtual/2025/poster/32053 abstract： Diffusion-based video generation technology has advanced significantly, catalyzing a proliferation of research in human animation. While breakthroughs have been made in driving human animation through various modalities for portraits, most of current solutions for human body animation still focus on video-driven methods, leaving audio-driven taking body generation relatively underexplored. In this paper, we introduce CyberHost, a one-stage audio-driven talking body generation framework that addresses common synthesis degradations in half-body animation, including hand integrity, identity consistency, and natural motion.CyberHost's key designs are twofold. Firstly, the Region Attention Module (RAM) maintains a set of learnable, implicit, identity-agnostic latent features and combines them with identity-specific local visual features to enhance the synthesis of critical local regions. Secondly, the Human-Prior-Guided Conditions introduce more human structural priors into the model, reducing uncertainty in generated motion patterns and thereby improving the stability of the generated videos.To our knowledge, CyberHost is the first one-stage audio-driven human diffusion model capable of zero-shot video generation for the human body. Extensive experiments demonstrate that CyberHost surpasses previous works in both quantitative and qualitative aspects. CyberHost can also be extended to video-driven and audio-video hybrid-driven scenarios, achieving similarly satisfactory results.

264. A Statistical Framework for Ranking LLM-based Chatbots

链接：https://iclr.cc/virtual/2025/poster/28215 abstract： Large language models (LLMs) have transformed natural language processing, with frameworks like Chatbot Arena providing pioneering platforms for evaluating these models. By facilitating millions of pairwise comparisons based on human judgments, Chatbot Arena has become a cornerstone in LLM evaluation, offering rich datasets for ranking models in open-ended conversational tasks. Building upon this foundation, we propose a statistical framework that incorporates key advancements to address specific challenges in pairwise comparison analysis. First, we introduce a factored tie model that enhances the ability to handle ties—an integral aspect of human-judged comparisons—significantly improving the model's fit to observed data. Second, we extend the framework to model covariance between competitors, enabling deeper insights into performance relationships and facilitating intuitive groupings into performance tiers. Third, we resolve optimization challenges arising from parameter non-uniqueness by introducing novel constraints, ensuring stable and interpretable parameter estimation. Through rigorous evaluation and extensive experimentation, our framework demonstrates substantial improvements over existing methods in modeling pairwise comparison data. To support reproducibility and practical adoption, we release leaderbot, an open-source Python package implementing our models and analyses.

265. KooNPro: A Variance-Aware Koopman Probabilistic Model Enhanced by Neural Process for Time Series Forecasting

链接：https://iclr.cc/virtual/2025/poster/30930 abstract： The probabilistic forecasting of time series is a well-recognized challenge, particularly in disentangling correlations among interacting time series and addressing the complexities of distribution modeling. By treating time series as temporal dynamics, we introduce KooNPro, a novel probabilistic time series forecasting model that combines variance-aware deep Koopman model with Neural Process. KooNPro introduces a variance-aware continuous spectrum using Gaussian distributions to capture complex temporal dynamics with improved stability. It further integrates the Neural Process to capture fine dynamics, enabling enhanced dynamics capture and prediction. Extensive experiments on nine real-world datasets demonstrate that KooNPro consistently outperforms state-of-the-art baselines. Ablation studies highlight the importance of the Neural Process component and explore the impact of key hyperparameters. Overall, KooNPro presents a promising novel approach for probabilistic time series forecasting.

266. Lie Algebra Canonicalization: Equivariant Neural Operators under arbitrary Lie Groups

链接：https://iclr.cc/virtual/2025/poster/30825 abstract： The quest for robust and generalizable machine learning models has driven recent interest in exploiting symmetries through equivariant neural networks. In the context of PDE solvers, recent works have shown that Lie point symmetries can be a useful inductive bias for Physics-Informed Neural Networks (PINNs) through data and loss augmentation. Despite this, directly enforcing equivariance within the model architecture for these problems remains

elusive. This is because many PDEs admit non-compact symmetry groups, oftentimes not studied beyond their infinitesimal generators, making them incompatible with most existing equivariant architectures. In this work, we propose Lie aLgebrA Canonicalization (LieLAC), a novel approach that exploits only the action of infinitesimal generators of the symmetry group, circumventing the need for knowledge of the full group structure. To achieve this, we address existing theoretical issues in the canonicalization literature, establishing connections with frame averaging in the case of continuous non-compact groups. Operating within the framework of canonicalization, LieLAC can easily be integrated with unconstrained pre-trained models, transforming inputs to a canonical form before feeding them into the existing model, effectively aligning the input for model inference according to allowed symmetries. LieLAC utilizes standard Lie group descent schemes, achieving equivariance in pre-trained models. Finally, we showcase LieLAC's efficacy on tasks of invariant image classification and Lie point symmetry equivariant neural PDE solvers using pre-trained models.

267. Block Verification Accelerates Speculative Decoding

链接：https://iclr.cc/virtual/2025/poster/28852 abstract： Speculative decoding is an effective method for lossless acceleration of large language models during inference. It uses a fast model to draft a block of tokens which are then verified in parallel by the target model, and provides a guarantee that the output is distributed identically to a sample from the target model. In prior works, draft verification is performed independently token-by-token. Surprisingly, we show that this approach is not optimal. We propose Block Verification, a simple draft verification algorithm that verifies the entire block jointly and provides additional wall-clock speedup. We prove that the proposed mechanism is optimal in the expected number of tokens produced each iteration and specifically is never worse than the standard token-level verification.Empirically, block verification provides modest but consistent wall-clock speedups over the standard token verification algorithm of 5\%-8\% in a range of tasks and datasets.Given that block verification does not increase code complexity, maintains the strong lossless guarantee of the standard speculative decoding verification algorithm, cannot deteriorate performance, and, in fact, consistently improves it, it can be used as a good default in speculative decoding implementations.

268. Continuous Diffusion for Mixed-Type Tabular Data

链接：https://iclr.cc/virtual/2025/poster/29690 abstract： Score-based generative models, commonly referred to as diffusion models, have proven to be successful at generating text and image data.However, their adaptation to mixed-type tabular data remains underexplored. In this work, we propose CDTD, a Continuous Diffusion model for mixed-type Tabular Data. CDTD is based on a novel combination of score matching and score interpolation to enforce a unified continuous noise distribution for both continuous and categorical features. We explicitly acknowledge the necessity of homogenizing distinct data types by relying on model-specific loss calibration and initialization schemes. To further address the high heterogeneity in mixed-type tabular data, we introduce adaptive feature- or type-specific noise schedules. These ensure balanced generative performance across features and optimize the allocation of model capacity across features and diffusion time. Our experimental results show that CDTD consistently outperforms state-of-the-art benchmark models, captures feature correlations exceptionally well, and that heterogeneity in the noise schedule design boosts sample quality. Replication code is available at https://github.com/muellermarkus/cdtd.

269. Intricacies of Feature Geometry in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/31349 abstract： Studying the geometry of a language model's embedding space is an important and challenging task because of the various ways concepts can be represented, extracted, and used. Specifically, we want a framework that unifies both measurement (of how well a latent explains a feature/concept) and causal intervention (how well it can be used to control/steer the model). We discuss several challenges with using some recent approaches to study the geometry of categorical and hierarchical concepts in large language models (LLMs) and both theoretically and empirically justify our main takeaway, which is that their orthogonality and polytopes results are trivially true in high-dimensional spaces, and can be observed even in settings where they should not occur.

270. What to align in multimodal contrastive learning?

链接：https://iclr.cc/virtual/2025/poster/29742 abstract： Humans perceive the world through multisensory integration, blending the information of different modalities to adapt their behavior.Contrastive learning offers an appealing solution for multimodal self-supervised learning. Indeed, by considering each modality as a different view of the same entity, it learns to align features of different modalities in a shared representation space. However, this approach is intrinsically limited as it only learns shared or redundant information between modalities, while multimodal interactions can arise in other ways. In this work, we introduce CoMM, a Contrastive Multimodal learning strategy that enables the communication between modalities in a single multimodal space. Instead of imposing cross- or intra- modality constraints, we propose to align multimodal representations by maximizing the mutual information between augmented versions of these multimodal features. Our theoretical analysis shows that shared, synergistic and unique terms of information naturally emerge from this formulation, allowing us to estimate multimodal interactions beyond redundancy. We test CoMM both in a controlled and in a series of real-world settings: in the former, we demonstrate that CoMM effectively captures redundant, unique and synergistic information between modalities. In the latter, CoMM learns complex multimodal interactions and achieves state-of-the-art results on seven multimodal tasks.

271. MIM-Refiner: A Contrastive Learning Boost from Intermediate Pre-

Trained Masked Image Modeling Representations

链接：https://iclr.cc/virtual/2025/poster/31256 abstract： We introduce MIM (Masked Image Modeling)-Refiner, a contrastive learning boost for pre-trained MIM models. MIM-Refiner is motivated by the insight that strong representations within MIM models generally reside in intermediate layers. Accordingly, MIM-Refiner leverages multiple instance discrimination (ID) heads that are connected to different intermediate layers. In each head, a nearest neighbor ID objective constructs clusters that capture semantic information which improves performance on downstream tasks, including off-the-shelf and fine-tuning settings.The refinement process is short and simple - yet highly effective. Within a few epochs, we refine the features of MIM models from subpar to state-of-the-art, off-the-shelf features. Refining a ViT-H, pre-trained with data2vec 2.0 on ImageNet-1K, sets a new state-of-the-art in linear probing (84.7\%) and low-shot classification among models that are pre-trained on ImageNet-1K. MIM-Refiner efficiently combines the advantages of MIM and ID objectives, enabling scaling ID objectives to billion parameter models using relatively little compute. MIM-Refiner compares favorably against previous state-of-the-art SSL models on various benchmarks such as low-shot classification, long-tailed classification and semantic segmentation.

272. Shape as Line Segments: Accurate and Flexible Implicit Surface Representation

链接：https://iclr.cc/virtual/2025/poster/29644 abstract： Distance field-based implicit representations like signed/unsigned distance fields have recently gained prominence in geometry modeling and analysis. However, these distance fields are reliant on the closest distance of points to the surface, introducing inaccuracies when interpolating along cube edges during surface extraction. Additionally, their gradients are ill-defined at certain locations, causing distortions in the extracted surfaces. To address this limitation, we propose Shape as Line Segments (SALS), an accurate and efficient implicit geometry representation based on attributed line segments, which can handle arbitrary structures. Unlike previous approaches, SALS leverages a differentiable Line Segment Field to implicitly capture the spatial relationship between line segments and the surface. Each line segment is associated with two key attributes, intersection flag and ratio, from which we propose edge-based dual contouring to extract a surface. We further implement SALS with a neural network, producing a new neural implicit presentation. Additionally, based on SALS, we design a novel learning-based pipeline for reconstructing surfaces from 3D point clouds. We conduct extensive experiments, showcasing the significant advantages of our methods over state-of-the-art methods.The source code is available at https://github.com/rsy6318/SALS.

273. Generative Flows on Synthetic Pathway for Drug Design

链接：https://iclr.cc/virtual/2025/poster/28313 abstract： Generative models in drug discovery have recently gained attention as efficient alternatives to brute-force virtual screening. However, most existing models do not account for synthesizability, limiting their practical use in real-world scenarios. In this paper, we propose RxnFlow, which sequentially assembles molecules using predefined molecular building blocks and chemical reaction templates to constrain the synthetic chemical pathway. We then train on this sequential generating process with the objective of generative flow networks (GFlowNets) to generate both highly rewarded and diverse molecules. To mitigate the large action space of synthetic pathways in GFlowNets, we implement a novel action space subsampling method. This enables RxnFlow to learn generative flows over extensive action spaces comprising combinations of 1.2 million building blocks and 71 reaction templates without significant computational overhead. Additionally, RxnFlow can employ modified or expanded action spaces for generation without retraining, allowing for the introduction of additional objectives or the incorporation of newly discovered building blocks. We experimentally demonstrate that RxnFlow outperforms existing reaction-based and fragment-based models in pocket-specific optimization across various target pockets. Furthermore, RxnFlow achieves state-of-the-art performance on CrossDocked2020 for pocket-conditional generation, with an average Vina score of –8.85 kcal/mol and 34.8% synthesizability. Code is available at https://github.com/SeonghwanSeo/RxnFlow.

274. DartControl: A Diffusion-Based Autoregressive Motion Model for Real-Time Text-Driven Motion Control

链接：https://iclr.cc/virtual/2025/poster/29308 abstract： Text-conditioned human motion generation, which allows for user interaction through natural language, has become increasingly popular. Existing methods typically generate short, isolated motions based on a single input sentence. However, human motions are continuous and can extend over long periods, carrying rich semantics. Creating long, complex motions that precisely respond to streams of text descriptions, particularly in an online and real-time setting, remains a significant challenge. Furthermore, incorporating spatial constraints into text-conditioned motion generation presents additional challenges, as it requires aligning the motion semantics specified by text descriptions with geometric information, such as goal locations and 3D scene geometry. To address these limitations, we propose DartControl, in short DART, a Diffusion-based Autoregressive motion primitive model for Real-time Text-driven motion Control. Our model, DART, effectively learns a compact motion primitive space jointly conditioned on motion history and text inputs using latent diffusion models. By autoregressively generating motion primitives based on the preceding history and current text input, DART enables real-time, sequential motion generation driven by natural language descriptions. Additionally, the learned motion primitive space allows for precise spatial motion control, which we formulate either as a latent noise optimization problem or as a Markov decision process addressed through reinforcement learning. We present effective algorithms for both approaches, demonstrating our model’s versatility and superior performance in various motion synthesis tasks. Experiments show our method outperforms existing baselines in motion realism, efficiency, and controllability. Video results and code are available at

https://zkf1997.github.io/DART/.

275. NExUME: Adaptive Training and Inference for DNNs under Intermittent Power Environments

链接：https://iclr.cc/virtual/2025/poster/29619 abstract： The deployment of Deep Neural Networks (DNNs) in energy-constrained environments, such as Energy Harvesting Wireless Sensor Networks (EH-WSNs), introduces significant challenges due to the intermittent nature of power availability. This study introduces NExUME, a novel training methodology designed specifically for DNNs operating under such constraints. We propose a dynamic adjustment of training parameters—dropout rates and quantization levels—that adapt in real-time to the available energy, which varies in energy harvesting scenarios.This approach utilizes a model that integrates the characteristics of the network architecture and the specific energy harvesting profile. It dynamically adjusts training strategies, such as the intensity and timing of dropout and quantization, based on predictions of energy availability. This method not only conserves energy but also enhances the network's adaptability, ensuring robust learning and inference capabilities even under stringent power constraints. Our results show a 6% to 22% improvement in accuracy over current methods, with an increase of less than 5% in computational overhead. This paper details the development of the adaptive training framework, describes the integration of energy profiles with dropout and quantization adjustments, and presents a comprehensive evaluation using real-world data. Additionally, we introduce a novel dataset aimed at furthering the application of energy harvesting in computational settings.

276. PEARL: Parallel Speculative Decoding with Adaptive Draft Length

链接：https://iclr.cc/virtual/2025/poster/29693 abstract： Speculative decoding (SD), where an extra draft model is employed to provide multiple **draft** tokens first and then the original target model verifies these tokens in parallel, has shown great power for LLM inference acceleration.However, existing SD methods suffer from the mutual waiting problem, i.e., the target model gets stuck when the draft model is *guessing* tokens, and vice versa. This problem is directly incurred by the asynchronous execution of the draft model and the target model, and is exacerbated due to the fixed draft length in speculative decoding.To address these challenges, we propose a conceptually simple, flexible, and general framework to boost speculative decoding, namely **P**arallel sp**E**culative decoding with **A**daptive d**R**aft **L**ength (PEARL). Specifically, PEARL proposes *pre-verify* to verify the first draft token in advance during the drafting phase, and *post-verify* to generate more draft tokens during the verification phase.PEARL parallels the drafting phase and the verification phase via applying the two strategies, and achieves adaptive draft length for different scenarios, which effectively alleviates the mutual waiting problem.Experiments on various text generation benchmarks demonstrate the effectiveness of our PEARL, leading to a superior speedup performance up to **4.43$\\times$** and **1.50$\\times$**, compared to auto-regressive decoding and vanilla speculative decoding, respectively.

277. Weakly Supervised Video Scene Graph Generation via Natural Language Supervision

链接：https://iclr.cc/virtual/2025/poster/30280 abstract： Existing Video Scene Graph Generation (VidSGG) studies are trained in a fully supervised manner, which requires all frames in a video to be annotated, thereby incurring high annotation cost compared to Image Scene Graph Generation (ImgSGG). Although the annotation cost of VidSGG can be alleviated by adopting a weakly supervised approach commonly used for ImgSGG (WS-ImgSGG) that uses image captions, there are two key reasons that hinder such a naive adoption: 1) Temporality within video captions, i.e., unlike image captions, video captions include temporal markers (e.g., before, while, then, after) that indicate time-related details, and 2) Variability in action duration, i.e., unlike human actions in image captions, human actions in video captions unfold over varying duration. To address these issues, we propose a Natural Language-based Video Scene Graph Generation (NL-VSGG) framework that only utilizes the readily available video captions for training a VidSGG model. NL-VSGG consists of two key modules: Temporality-aware Caption Segmentation (TCS) module and Action Duration Variability-aware caption-frame alignment (ADV) module. Specifically, TCS segments the video captions into multiple sentences in a temporal order based on a Large Language Model (LLM), and ADV aligns each segmented sentence with appropriate frames considering the variability in action duration. Our approach leads to a significant enhancement in performance compared to simply applying the WS-ImgSGG pipeline to VidSGG on the Action Genome dataset. As a further benefit of utilizing the video captions as weak supervision, we show that the VidSGG model trained by NL-VSGG is able to predict a broader range of action classes that are not included in the training data, which makes our framework practical in reality.

278. Mixture of Attentions For Speculative Decoding

链接：https://iclr.cc/virtual/2025/poster/29634 abstract： The growth in the number of parameters of Large Language Models (LLMs) has led to a significant surge in computational requirements, making them challenging and costly to deploy.Speculative decoding (SD) leverages smaller models to efficiently propose future tokens, which are then verified by the LLM in parallel.Small models that utilise activations from the LLM currently achieve the fastest decoding speeds.However, we identify several limitations of SD models including the lack of on-policyness during training and partial observability. To address these shortcomings, we propose a more grounded architecture for small models by introducing a Mixture of Attentions for SD.Our novel architecture can be applied in two scenarios: a conventional single device deployment and a novel client-server deployment where the small model is hosted on a consumer device and the LLM on a server.In a single-device scenario, we demonstrate state-of-the-art speedups improving EAGLE-2 by 9.5% and its acceptance length by 25%.In a client-server setting,

our experiments demonstrate: 1) state-of-the-art latencies with minimal calls to the server for different network conditions, and 2) in the event of a complete disconnection, our approach can maintain higher accuracy compared to other SD methods and demonstrates advantages over API calls to LLMs, which would otherwise be unable to continue the generation process.

279. Data Shapley in One Training Run

链接：https://iclr.cc/virtual/2025/poster/30239 abstract： Data Shapley offers a principled framework for attributing the contribution of data within machine learning contexts. However, the traditional notion of Data Shapley requires re-training models on various data subsets, which becomes computationally infeasible for large-scale models. Additionally, this retraining-based definition cannot evaluate the contribution of data for a specific model training run, which may often be of interest in practice. This paper introduces a novel concept, In-Run Data Shapley, which eliminates the need for model retraining and is specifically designed for assessing data contribution for a particular model of interest. In-Run Data Shapley calculates the Shapley value for each gradient update iteration and accumulates these values throughout the training process. We present several techniques that allow the efficient scaling of In-Run Data Shapley to the size of foundation models. In its most optimized implementation, our method adds negligible runtime overhead compared to standard model training. This dramatic efficiency improvement makes it possible to perform data attribution for the foundation model pretraining stage. We present several case studies that offer fresh insights into pretraining data's contribution and discuss their implications for copyright in generative AI and pretraining data curation.

280. PALMBENCH: A COMPREHENSIVE BENCHMARK OF COMPRESSED LARGE LANGUAGE MODELS ON MOBILE PLATFORMS

链接：https://iclr.cc/virtual/2025/poster/27755 abstract： Deploying large language models (LLMs) locally on mobile devices is advantageous in scenarios where transmitting data to remote cloud servers is either undesirable due to privacy concerns or impractical due to network connection. Recent advancements have facilitated the local deployment of LLMs. However, local deployment also presents challenges, particularly in balancing quality (generative performance), latency, and throughput within the hardware constraints of mobile devices. In this paper, we introduce our lightweight, all-in-one automated benchmarking framework that allows users to evaluate LLMs on mobile devices. We provide a comprehensive benchmark of various popular LLMs with different quantization configurations (both weights and activations) across multiple mobile platforms with varying hardware capabilities. Unlike traditional benchmarks that assess full-scale models on high-end GPU clusters, we focus on evaluating resource efficiency (memory and power consumption) and harmful output for compressed models on mobile devices. Our key observations include: i) differences in energy efficiency and throughput across mobile platforms; ii) the impact of quantization on memory usage, GPU execution time, and power consumption; and iii) accuracy and performance degradation of quantized models compared to their non-quantized counterparts; and iv) the frequency of hallucinations and toxic content generated by compressed LLMs onmobile devices.

281. Provably Robust Explainable Graph Neural Networks against Graph Perturbation Attacks

链接：https://iclr.cc/virtual/2025/poster/28705 abstract： Explaining Graph Neural Network (XGNN) has gained growing attention to facilitate the trust of using GNNs, which is the mainstream method to learn graph data. Despite their growing attention, Existing XGNNs focus on improving the explanation performance, and its robustness under attacks is largely unexplored. We noticed that an adversary can slightly perturb the graph structure such that the explanation result of XGNNs is largely changed. Such vulnerability of XGNNs could cause serious issues particularly in safety/security-critical applications. In this paper, we take the first step to study the robustness of XGNN against graph perturbation attacks, and propose XGNNCert, the first provably robust XGNN. Particularly, our XGNNCert can provably ensure the explanation result for a graph under the worst-case graph perturbation attack is close to that without the attack, while not affecting the GNN prediction, when the number of perturbed edges is bounded. Evaluation results on multiple graph datasets and GNN explainers show the effectiveness of XGNNCert.

282. What Does It Mean to Be a Transformer? Insights from a Theoretical Hessian Analysis

链接：https://iclr.cc/virtual/2025/poster/31064 abstract： The Transformer architecture has inarguably revolutionized deep learning, overtaking classical architectures like multi-layer perceptions (MLPs) and convolutional neural networks (CNNs). At its core, the attention block differs in form and functionality from most other architectural components in deep learning—to the extent that, in comparison to MLPs/CNNs, Transformers are more often accompanied by adaptive optimizers, layer normalization, learning rate warmup, etc. The root causes behind these outward manifestations and the precise mechanisms that govern them remain poorly understood. In this work, we bridge this gap by providing a fundamental understanding of what distinguishes the Transformer from the other architectures—grounded in a theoretical comparison of the (loss) Hessian. Concretely, for a single self-attention layer, (a) we first entirely derive the Transformer’s Hessian and express it in matrix derivatives; (b) we then characterize it in terms of data, weight, and attention moment dependencies; and (c) while doing so further highlight the important structural differences to the Hessian of classical networks. Our results suggest that various common architectural and optimization choices in Transformers can be traced back to their highly non-linear dependencies on the data and weight

matrices, which vary heterogeneously across parameters. Ultimately, our findings provide a deeper understanding of the Transformer’s unique optimization landscape and the challenges it poses.

283. Can Reinforcement Learning Solve Asymmetric Combinatorial-Continuous Zero-Sum Games?

链接：https://iclr.cc/virtual/2025/poster/30815 abstract： There have been extensive studies on learning in zero-sum games, focusing on the analysis of the existence and algorithmic convergence of Nash equilibrium (NE). Existing studies mainly focus on symmetric games where the strategy spaces of the players are of the same type and size. For the few studies that do consider asymmetric games, they are mostly restricted to matrix games. In this paper, we define and study a new practical class of asymmetric games called two-player Asymmetric Combinatorial-Continuous zEro-Sum (ACCES) games, featuring a combinatorial action space for one player and an infinite compact space for the other. Such ACCES games have broad implications in the real world, particularly in combinatorial optimization problems (COPs) where one player optimizes a solution in a combinatorial space, and the opponent plays against it in an infinite (continuous) compact space (e.g., a nature player deciding epistemic parameters of the environmental model). Our first key contribution is to prove the existence of NE for two-player ACCES games, using the idea of essentially finite game approximation. Building on the theoretical insights and double oracle (DO)-based solutions to complex zero-sum games, our second contribution is to design the novel algorithm, Combinatorial Continuous DO (CCDO), to solve ACCES games, and prove the convergence of the proposed algorithm. Considering the NP-hardness of most COPs and recent advancements in reinforcement learning (RL)-based solutions to COPs, our third contribution is to propose a practical algorithm to solve NE in the real world, CCDORL (based on CCDO) and provide the novel convergence analysis in the ACCES game. Experimental results across diverse instances of COPs demonstrate the empirical effectiveness of our algorithms.

284. Beyond Linear Approximations: A Novel Pruning Approach for Attention Matrix

链接：https://iclr.cc/virtual/2025/poster/28105 abstract： Large Language Models (LLMs) have shown immense potential in enhancing various aspects of our daily lives, from conversational AI to search and AI assistants. However, their growing capabilities come at the cost of extremely large model sizes, making deployment on edge devices challenging due to memory and computational constraints. This paper introduces a novel approach to LLM weight pruning that directly optimizes for approximating the attention matrix, a core component of transformer architectures. Unlike existing methods that focus on linear approximations, our approach accounts for the non-linear nature of the Softmax attention mechanism. We provide theoretical guarantees for the convergence of our Gradient Descent-based optimization method to a near-optimal pruning mask solution. Our empirical results demonstrate the effectiveness of our non-linear pruning approach in maintaining model performance while significantly reducing computational costs, which is beyond the current state-of-the-art methods, i.e., SparseGPT and Wanda, by a large margin. This work establishes a new theoretical foundation for pruning algorithm design in LLMs, potentially paving the way for more efficient LLM inference on resource-constrained devices.

285. Standardizing Structural Causal Models

链接：https://iclr.cc/virtual/2025/poster/29160 abstract： Synthetic datasets generated by structural causal models (SCMs) are commonly used for benchmarking causal structure learning algorithms. However, the variances and pairwise correlations in SCM data tend to increase along the causal ordering. Several popular algorithms exploit these artifacts, possibly leading to conclusions that do not generalize to real-world settings. Existing metrics like Var-sortability and $\operatorname{R^2}$-sortability quantify these patterns, but they do not provide tools to remedy them. To address this, we propose internally-standardized structural causal models (iSCMs), a modification of SCMs that introduces a standardization operation at each variable during the generative process. By construction, iSCMs are not Var-sortable. We also find empirical evidence that they are mostly not $\operatorname{R^2}$-sortable for commonly-used graph families. Moreover, contrary to the post-hoc standardization of data generated by standard SCMs, we prove that linear iSCMs are less identifiable from prior knowledge on the weights and do not collapse to deterministic relationships in large systems, which may make iSCMs a useful model in causal inference beyond the benchmarking problem studied here. Our code is publicly available at: https://github.com/werkaaa/iscm.

286. Capturing the Temporal Dependence of Training Data Influence

链接：https://iclr.cc/virtual/2025/poster/27988 abstract： Traditional data influence estimation methods, like influence function, assume that learning algorithms are permutation-invariant with respect to training data. However, modern training paradigms—especially for foundation models using stochastic algorithms and non-convergent, multi-stage curricula—are sensitive to data ordering, thus violating this assumption. This mismatch renders influence functions inadequate for answering some critical questions in current machine learning: How can we differentiate the influence of the same data contributing at different stages of training? More generally, how can we capture the dependence of data influence on the optimization trajectory during training? To address this gap, we formalize the concept of \emph{trajectory-specific leave-one-out (LOO) influence}, which quantifies the impact of removing a data point from a specific iteration during training, accounting for the exact sequence of data encountered and the model's optimization trajectory. However, exactly evaluating the trajectory-specific LOO presents a significant computational challenge. To address this, we propose \emph{data value embedding}, a novel technique enabling efficient

approximation of trajectory-specific LOO. Specifically, we compute a training data embedding that encapsulates the cumulative interactions between data and the evolving model parameters. The LOO can then be efficiently approximated through a simple dot-product between the data value embedding and the gradient of the given test data. As data value embedding captures training data ordering, it offers valuable insights into model training dynamics. In particular, we uncover distinct phases of data influence, revealing that data points in the early and late stages of training exert a greater impact on the final model. These insights translate into actionable strategies for managing the computational overhead of data selection by strategically timing the selection process, potentially opening new avenues in data curation research.

287. Reinforcement learning with combinatorial actions for coupled restless bandits

链接：https://iclr.cc/virtual/2025/poster/30448 abstract： Reinforcement learning (RL) has increasingly been applied to solve real-world planning problems, with progress in handling large state spaces and time horizons. However, a key bottleneck in many domains is that RL methods cannot accommodate large, combinatorially structured action spaces. In such settings, even representing the set of feasible actions at a single step may require a complex discrete optimization formulation. We leverage recent advances in embedding trained neural networks into optimization problems to propose SEQUOIA, an RL algorithm that directly optimizes for long-term reward over the feasible action space. Our approach embeds a Q-network into a mixed-integer program to select a combinatorial action in each timestep. Here, we focus on planning over restless bandits, a class of planning problems which capture many real-world examples of sequential decision making. We introduce coRMAB, a broader class of restless bandits with combinatorial actions that cannot be decoupled across the arms of the restless bandit, requiring direct solving over the joint, exponentially large action space. We empirically validate SEQUOIA on four novel restless bandit problems with combinatorial constraints: multiple interventions, path constraints, bipartite matching, and capacity constraints. Our approach significantly outperforms existing methods—which cannot address sequential planning and combinatorial selection simultaneously—by an average of 24.8% on these difficult instances.

288. Diffusion Bridge AutoEncoders for Unsupervised Representation Learning

链接：https://iclr.cc/virtual/2025/poster/28774 abstract： Diffusion-based representation learning has achieved substantial attention due to its promising capabilities in latent representation and sample generation. Recent studies have employed an auxiliary encoder to identify a corresponding representation from data and to adjust the dimensionality of a latent variable $\mathbf{z}$. Meanwhile, this auxiliary structure invokes an *information split problem*; the information of each data instance $\mathbf{x}_0$ is divided into diffusion endpoint $\mathbf{x}_T$ and encoded $\mathbf{z}$ because there exist two inference paths starting from the data. The latent variable modeled by diffusion endpoint $\mathbf{x}_T$ has some disadvantages. The diffusion endpoint $\mathbf{x}_T$ is computationally expensive to obtain and inflexible in dimensionality. To address this problem, we introduce Diffusion Bridge AuteEncoders (DBAE), which enables $\mathbf{z}$-dependent endpoint $\mathbf{x}_T$ inference through a feed-forward architecture. This structure creates an information bottleneck at $\mathbf{z}$, so $\mathbf{x}_T$ becomes dependent on $\mathbf{z}$ in its generation. This results in $\mathbf{z}$ holding the full information of data. We propose an objective function for DBAE to enable both reconstruction and generative modeling, with their theoretical justification. Empirical evidence supports the effectiveness of the intended design in DBAE, which notably enhances downstream inference quality, reconstruction, and disentanglement. Additionally, DBAE generates high-fidelity samples in the unconditional generation. Our code isavailable at https://github.com/aailab-kaist/DBAE.

289. CL-MFAP: A Contrastive Learning-Based Multimodal Foundation Model for Molecular Property Prediction and Antibiotic Screening

链接：https://iclr.cc/virtual/2025/poster/28842 abstract： Due to the rise in antimicrobial resistance, identifying novel compounds with antibiotic potential is crucial for combatting this global health issue. However, traditional drug development methods are costly and inefficient. Recognizing the pressing need for more effective solutions, researchers have turned to machine learning techniques to streamline the prediction and development of novel antibiotic compounds. While foundation models have shown promise in antibiotic discovery, current mainstream efforts still fall short of fully leveraging the potential of multimodal molecular data. Recent studies suggest that contrastive learning frameworks utilizing multimodal data exhibit excellent performance in representation learning across various domains. Building upon this, we introduce CL-MFAP, an unsupervised contrastive learning (CL)-based multimodal foundation (MF) model specifically tailored for discovering small molecules with potential antibiotic properties (AP) using three types of molecular data. This model employs 1.6 million bioactive molecules with drug-like properties from the ChEMBL dataset to jointly pretrain three encoders: (1) a transformer-based encoder with rotary position embedding for processing SMILES strings; (2) another transformer-based encoder, incorporating a novel bi-level routing attention mechanism to handle molecular graph representations; and (3) a Morgan fingerprint encoder using a multilayer perceptron, to achieve the contrastive learning purpose. The CL-MFAP outperforms baseline models in antibiotic property prediction by effectively utilizing different molecular modalities and demonstrates superior domain-specific performance when fine-tuned for antibiotic-related property prediction tasks.

290. Perm: A Parametric Representation for Multi-Style 3D Hair Modeling

链接：https://iclr.cc/virtual/2025/poster/29369 abstract： We present Perm, a learned parametric representation of human 3D hair designed to facilitate various hair-related applications. Unlike previous work that jointly models the global hair structure and local curl patterns, we propose to disentangle them using a PCA-based strand representation in the frequency domain, thereby allowing more precise editing and output control. Specifically, we leverage our strand representation to fit and decompose hair geometry textures into low- to high-frequency hair structures, termed guide textures and residual textures, respectively. These decomposed textures are later parameterized with different generative models, emulating common stages in the hair grooming process. We conduct extensive experiments to validate the architecture design of Perm, and finally deploy the trained model as a generic prior to solve task-agnostic problems, further showcasing its flexibility and superiority in tasks such as single-view hair reconstruction, hairstyle editing, and hair-conditioned image generation. More details can be found on our project page: https://cs.yale.edu/homes/che/projects/perm/.

291. Reading Your Heart: Learning ECG Words and Sentences via Pre-training ECG Language Model

链接：https://iclr.cc/virtual/2025/poster/30897 abstract： Electrocardiogram (ECG) is essential for the clinical diagnosis of arrhythmias and other heart diseases, but deep learning methods based on ECG often face limitations due to the need for high-quality annotations. Although previous ECG self-supervised learning (eSSL) methods have made significant progress in representation learning from unannotated ECG data, they typically treat ECG signals as ordinary time-series data, segmenting the signals using fixed-size and fixed-step time windows, which often ignore the form and rhythm characteristics and latent semantic relationships in ECG signals. In this work, we introduce a novel perspective on ECG signals, treating heartbeats as words and rhythms as sentences. Based on this perspective, we first designed the QRS-Tokenizer, which generates semantically meaningful ECG sentences from the raw ECG signals. Building on these, we then propose HeartLang, a novel self-supervised learning framework for ECG language processing, learning general representations at form and rhythm levels. Additionally, we construct the largest heartbeat-based ECG vocabulary to date, which will further advance the development of ECG language processing. We evaluated HeartLang across six public ECG datasets, where it demonstrated robust competitiveness against other eSSL methods. Our data and code are publicly available at https://github.com/PKUDigitalHealth/HeartLang.

292. Accelerating Goal-Conditioned Reinforcement Learning Algorithms and Research

链接：https://iclr.cc/virtual/2025/poster/31000 abstract： Self-supervision has the potential to transform reinforcement learning (RL), paralleling the breakthroughs it has enabled in other areas of machine learning. While self-supervised learning in other domains aims to find patterns in a fixed dataset, self-supervised goal-conditioned reinforcement learning (GCRL) agents discover *new* behaviors by learning from the goals achieved during unstructured interaction with the environment. However, these methods have failed to see similar success, both due to a lack of data from slow environment simulations as well as a lack of stable algorithms. We take a step toward addressing both of these issues by releasing a high-performance codebase and benchmark (`JaxGCRL`) for self-supervised GCRL, enabling researchers to train agents for millions of environment steps in minutes on a single GPU. By utilizing GPU-accelerated replay buffers, environments, and a stable contrastive RL algorithm, we reduce training time by up to \$22\times\$. Additionally, we assess key design choices in contrastive RL, identifying those that most effectively stabilize and enhance training performance. With this approach, we provide a foundation for future research in self-supervised GCRL, enabling researchers to quickly iterate on new ideas and evaluate them in diverse and challenging environments. Code: https://anonymous.4open.science/r/JaxGCRL-2316/README.md

293. A Decade's Battle on Dataset Bias: Are We There Yet?

链接：https://iclr.cc/virtual/2025/poster/29591 abstract： We revisit the ``dataset classification'' experiment suggested by Torralba & Efros (2011) a decade ago, in the new era with large-scale, diverse, and hopefully less biased datasets as well as more capable neural network architectures. Surprisingly, we observe that modern neural networks can achieve excellent accuracy in classifying which dataset an image is from: e.g., we report 84.7% accuracy on held-out validation data for the three-way classification problem consisting of the YFCC, CC, and DataComp datasets. Our further experiments show that such a dataset classifier could learn semantic features that are generalizable and transferable, which cannot be explained by memorization. We hope our discovery will inspire the community to rethink issues involving dataset bias.

294. An Effective Theory of Bias Amplification

链接：https://iclr.cc/virtual/2025/poster/29400 abstract： Machine learning models can capture and amplify biases present in data, leading to disparate test performance across social groups. To better understand, evaluate, and mitigate these biases, a deeper theoretical understanding of how model design choices and data distribution properties contribute to bias is needed. In this work, we contribute a precise analytical theory in the context of ridge regression, both with and without random projections, where the former models feedforward neural networks in a simplified regime. Our theory offers a unified and rigorous explanation of machine learning bias, providing insights into phenomena such as bias amplification and minority-group bias in various feature and parameter regimes. For example, we observe that there may be an optimal regularization penalty or training time to avoid bias amplification, and there can be differences in test error between groups that are not alleviated with increased parameterization. Importantly, our theoretical predictions align with empirical observations reported in the literature on machine

learning bias. We extensively empirically validate our theory on synthetic and semi-synthetic datasets.

295. Warm Diffusion: Recipe for Blur-Noise Mixture Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/28184 abstract： Diffusion probabilistic models have achieved remarkable success in generative tasks across diverse data types. While recent studies have explored alternative degradation processes beyond Gaussian noise, this paper bridges two key diffusion paradigms: hot diffusion, which relies entirely on noise, and cold diffusion, which uses only blurring without noise. We argue that hot diffusion fails to exploit the strong correlation between high-frequency image detail and low-frequency structures, leading to random behaviors in the early steps of generation. Conversely, while cold diffusion leverages image correlations for prediction, it neglects the role of noise (randomness) in shaping the data manifold, resulting in out-of-manifold issues and partially explaining its performance drop. To integrate both strengths, we propose Warm Diffusion, a unified Blur-Noise Mixture Diffusion Model (BNMD), to control blurring and noise jointly. Our divide-and-conquer strategy exploits the spectral dependency in images, simplifying score model estimation by disentangling the denoising and deblurring processes. We further analyze the Blur-to-Noise Ratio (BNR) using spectral analysis to investigate the trade-off between model learning dynamics and changes in the data manifold. Extensive experiments across benchmarks validate the effectiveness of our approach for image generation.

296. BRIGHT: A Realistic and Challenging Benchmark for Reasoning-Intensive Retrieval

链接：https://iclr.cc/virtual/2025/poster/27702 abstract： Existing retrieval benchmarks primarily consist of information-seeking queries (e.g., aggregated questions from search engines) where keyword or semantic-based retrieval is usually sufficient. However, many complex real-world queries require in-depth reasoning to identify relevant documents that go beyond surface form matching. For example, finding documentation for a coding question requires understanding the logic and syntax of the functions involved. To better benchmark retrieval on such challenging queries, we introduce BRIGHT, the first text retrieval benchmark that requires intensive reasoning to retrieve relevant documents. Our dataset consists of 1,398 real-world queries spanning diverse domains such as economics, psychology, mathematics, coding, and more. These queries are drawn from naturally occurring or carefully curated human data. Extensive evaluation reveals that even state-of-the-art retrieval models perform poorly on BRIGHT. The leading model on the MTEB leaderboard (Muennighoff et al., 2023), which achieves a score of 59.0 nDCG@10,1 produces a score of nDCG@10 of 18.0 on BRIGHT. We show that incorporating explicit reasoning about the query improves retrieval performance by up to 12.2 points. Moreover, incorporating retrieved documents from the top-performing retriever boosts question answering performance by over 6.6 points. We believe that BRIGHT paves the way for future research on retrieval systems in more realistic and challenging settings.

297. Reconciling Model Multiplicity for Downstream Decision Making

链接：https://iclr.cc/virtual/2025/poster/27942 abstract： We consider the problem of model multiplicity in downstream decision-making, a setting where two predictive models of equivalent accuracy cannot agree on what action to take for a downstream decision-making problem. Prior work attempts to address model multiplicity by resolving prediction disagreement between models. However, we show that even when the two predictive models approximately agree on their individual predictions almost everywhere, these models can lead the downstream decision-maker to take actions with substantially higher losses. We address this issue by proposing a framework that calibrates the predictive models with respect to both a finite set of downstream decision-making problems and the individual probability prediction. Specifically, leveraging tools from multi-calibration, we provide an algorithm that, at each time-step, first reconciles the differences in individual probability prediction, then calibrates the updated models such that they are indistinguishable from the true probability distribution to the decision-makers. We extend our results to the setting where one does not have direct access to the true probability distribution and instead relies on a set of i.i.d data to be the empirical distribution. Furthermore, we generalize our results to the settings where one has more than two predictive models and an infinitely large downstream action set. Finally, we provide a set of experiments to evaluate our methods empirically. Compared to existing work, our proposed algorithm creates a pair of predictive models with improved downstream decision-making losses and agrees on their best-response actions almost everywhere.

298. Unlearning or Obfuscating? Jogging the Memory of Unlearned LLMs via Benign Relearning

链接：https://iclr.cc/virtual/2025/poster/28883 abstract： Machine unlearning is a promising approach to mitigate undesirable memorization of training data in ML models. However, in this work we show that existing approaches for unlearning in LLMs are surprisingly susceptible to a simple set of benign relearning attacks. With access to only a small and potentially loosely related set of data, we find that we can “jog” the memory of unlearned models to reverse the effects of unlearning. For example, we show that relearning on public medical articles can lead an unlearned LLM to output harmful knowledge about bioweapons, and relearning general wiki information about the book series Harry Potter can force the model to output verbatim memorized text. We formalize this unlearning-relearning pipeline, explore the attack across three popular unlearning benchmarks, and discuss future directions and guidelines that result from our study. Our work indicates that current approximate unlearning methods simply suppress the model outputs and fail to robustly forget target knowledge in the LLMs.

299. ChroKnowledge: Unveiling Chronological Knowledge of Language Models in Multiple Domains

链接：https://iclr.cc/virtual/2025/poster/27828 abstract： Large language models (LLMs) have brought significant changes to many aspects of our lives.However, assessing and ensuring their chronological knowledge remains challenging.Existing approaches fall short in addressing the temporal adaptability of knowledge, often relying on a fixed time-point view. To overcome this, we introduce ChroKnowBench, a benchmark dataset designed to evaluate chronologically accumulated knowledge across three key aspects: multiple domains, time dependency, temporal state.Our benchmark distinguishes between knowledge that evolves (e.g., personal history, scientific discoveries, amended laws) and knowledge that remain constant (e.g., mathematical truths, commonsense facts). Building on this benchmark, we present ChroKnowledge (Chronological Categorization of Knowledge), a novel sampling-based framework for evaluating LLMs' non-parametric chronological knowledge.Our evaluation led to the following observations: (1) The ability of eliciting temporal knowledge varies depending on the data format that model was trained on.(2) LLMs partially recall knowledge or show a cut-off at temporal boundaries rather than recalling all aspects of knowledge correctly.Thus, we apply our ChroKnowPrompt, an in-depth prompting to elicit chronological knowledge by traversing step-by-step through the surrounding time spans.We observe that it successfully recalls objects across both open-source and proprietary LLMs, demonstrating versatility, though it faces challenges with dynamic datasets and unstructured formats.

300. Rethinking Evaluation of Sparse Autoencoders through the Representation of Polysemous Words

链接：https://iclr.cc/virtual/2025/poster/30209 abstract： Sparse autoencoders (SAEs) have gained a lot of attention as a promising tool to improve the interpretability of large language models (LLMs) by mapping the complex superposition of *polysemantic* neurons into *monosemantic* features and composing a sparse dictionary of words.However, traditional performance metrics like Mean Squared Error and L_{0} sparsity ignore the evaluation of the semantic representational power of SAEs - whether they can acquire interpretable monosemantic features while preserving the semantic relationship of words.For instance, it is not obvious whether a learned sparse feature could distinguish different meanings in one word.In this paper, we propose a suite of evaluations for SAEs to analyze the quality of monosemantic features by focusing on polysemous words.Our findings reveal that SAEs developed to improve the MSE-L_0 Pareto frontier may confuse interpretability, which does not necessarily enhance the extraction of monosemantic features.The analysis of SAEs with polysemous words can also figure out the internal mechanism of LLMs; deeper layers and the Attention module contribute to distinguishing polysemy in a word.Our semantics-focused evaluation offers new insights into the polysemy and the existing SAE objective and contributes to the development of more practical SAEs.

301. Procedural Knowledge in Pretraining Drives Reasoning in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/31190 abstract： The capabilities and limitations of Large Language Models (LLMs) have been sketched out in great detail in recent years, providing an intriguing yet conflicting picture. On the one hand, LLMs demonstrate a general ability to solve problems. On the other hand, they show surprising reasoning gaps when compared to humans, casting doubt on the robustness of their generalisation strategies. The sheer volume of data used in the design of LLMs has precluded us from applying the method traditionally used to measure generalisation: train-test set separation. To overcome this, we study what kind of generalisation strategies LLMs employ when performing reasoning tasks by investigating the pretraining data they rely on. For two models of different sizes (7B and 35B) and 2.5B of their pretraining tokens, we identify what documents influence the model outputs for three simple mathematical reasoning tasks and contrast this to the data that are influential for answering factual questions. We find that, while the models rely on mostly distinct sets of data for each factual question, a document often has a similar influence across different reasoning questions within the same task, indicating the presence of procedural knowledge. We further find that the answers to factual questions often show up in the most influential data. However, for reasoning questions the answers usually do not show up as highly influential, nor do the answers to the intermediate reasoning steps. When we characterise the top ranked documents for the reasoning questions qualitatively, we confirm that the influential documents often contain procedural knowledge, like demonstrating how to obtain a solution using formulae or code. Our findings indicate that the approach to reasoning the models use is unlike retrieval, and more like a generalisable strategy that synthesises procedural knowledge from documents doing a similar form of reasoning.

302. Calibrating Expressions of Certainty

链接：https://iclr.cc/virtual/2025/poster/28989 abstract： We present a novel approach to calibrating linguistic expressions of certainty, e.g., "Maybe" and "Likely". Unlike prior work that assigns a single score to each certainty phrase, we model uncertainty as distributions over the simplex to capture their semantics more accurately. To accommodate this new representation of certainty, we generalize existing measures of miscalibration and introduce a novel post-hoc calibration method. Leveraging these tools, we analyze the calibration of both humans (e.g., radiologists) and computational models (e.g., language models) and provide interpretable suggestions to improve their calibration.

303. Almost Optimal Batch-Regret Tradeoff for Batch Linear Contextual

Bandits

链接：https://iclr.cc/virtual/2025/poster/28187 abstract： We study the optimal batch-regret tradeoff for batch linear contextual bandits. For this problem, we design batch learning algorithms and prove that they achieve the optimal regret bounds (up to logarithmic factors) for any batch number M, number of actions K, time horizon T, and dimension d. Therefore, we establish the \emph{full-parameter-range} (almost) optimal batch-regret tradeoff for the batch linear contextual bandit problem. Along our analysis, we also prove a new matrix concentration inequality with dependence on their dynamic upper bounds, which, to the best of our knowledge, is the first of its kind in literature and maybe of independent interest.

304. Convergent Privacy Loss of Noisy-SGD without Convexity and Smoothness

链接：https://iclr.cc/virtual/2025/poster/28563 abstract： We study the Differential Privacy (DP) guarantee of hidden-state Noisy-SGD algorithms over a bounded domain. Standard privacy analysis for Noisy-SGD assumes all internal states are revealed, which leads to a divergent R\'enyi DP bound with respect to the number of iterations. Ye & Shokri (2022) and Altschuler & Talwar (2022) proved convergent bounds for smooth (strongly) convex losses, and raise open questions about whether these assumptions can be relaxed. We provide positive answers by proving convergent R\'enyi DP bound for non-convex non-smooth losses, where we show that requiring losses to have H\"older continuous gradient is sufficient. We also provide a strictly better privacy bound compared to state-of-the-art results for smooth strongly convex losses. Our analysis relies on the improvement of shifted divergence analysis in multiple aspects, including forward Wasserstein distance tracking, identifying the optimal shifts allocation, and the H\"older reduction lemma. Our results further elucidate the benefit of hidden-state analysis for DP and its applicability.

305. GLOMA: Global Video Text Spotting with Morphological Association

链接：https://iclr.cc/virtual/2025/poster/28058 abstract： Video Text Spotting (VTS) is a fundamental visual task that aims to predict the trajectories and content of texts in a video. Previous works usually conduct local associations and apply IoU-based distance and complex post-processing procedures to boost performance, ignoring the abundant temporal information and the morphological characteristics in VTS. In this paper, we propose \model{} to model the tracking problem as global associations and utilize the Gaussian Wasserstein distance to guide the morphological correlation between frames. Our main contributions can be summarized as three folds. 1). We propose a Transformer-based global tracking method \model{} for VTS and associate multiple frames simultaneously. 2). We introduce a Wasserstein distance-based method to conduct positional associations between frames. 3). We conduct extensive experiments on public datasets. On the ICDAR2015 video dataset, \model{} achieves \textbf{56.0} MOTA with \textbf{4.6} absolute improvement compared with the previous SOTA method and outperforms the previous Transformer-based method by a significant \textbf{8.3} MOTA.

306. Denoising as Adaptation: Noise-Space Domain Adaptation for Image Restoration

链接：https://iclr.cc/virtual/2025/poster/28613 abstract： Although learning-based image restoration methods have made significant progress, they still struggle with limited generalization to real-world scenarios due to the substantial domain gap caused by training on synthetic data. Existing methods address this issue by improving data synthesis pipelines, estimating degradation kernels, employing deep internal learning, and performing domain adaptation and regularization. Previous domain adaptation methods have sought to bridge the domain gap by learning domain-invariant knowledge in either feature or pixel space. However, these techniques often struggle to extend to low-level vision tasks within a stable and compact framework. In this paper, we show that it is possible to perform domain adaptation via the noise space using diffusion models. In particular, by leveraging the unique property of how auxiliary conditional inputs influence the multi-step denoising process, we derive a meaningful diffusion loss that guides the restoration model in progressively aligning both restored synthetic and real-world outputs with a target clean distribution. We refer to this method as denoising as adaptation. To prevent shortcuts during joint training, we present crucial strategies such as channel-shuffling layer and residual-swapping contrastive learning in the diffusion model. They implicitly blur the boundaries between conditioned synthetic and real data and prevent the reliance of the model on easily distinguishable features. Experimental results on three classical image restoration tasks, namely denoising, deblurring, and deraining, demonstrate the effectiveness of the proposed method.

307. MR-GSM8K: A Meta-Reasoning Benchmark for Large Language Model Evaluation

链接：https://iclr.cc/virtual/2025/poster/29086 abstract： In this work, we introduce a novel evaluation paradigm for Large Language Models(LLMs) that compels them to transition from a traditional question-answering role,akin to a student, to a solution-scoring role, akin to a teacher. This paradigm, focusing on "reasoning about reasoning," termed meta-reasoning, shifts the emphasisfrom result-oriented assessments, which often neglect the reasoning process, to amore comprehensive evaluation that effectively distinguishes between the cognitivecapabilities of different models. Our meta-reasoning process mirrors "system-2"slow thinking, requiring careful examination of assumptions, conditions, calculations, and logic to identify mistakes. This paradigm enables one to transformexisted saturated, non-differentiating benchmarks that might be leaked in data

pretraining stage to evaluation tools that are both challenging and robust against datacontamination. To prove our point, we applied our paradigm to GSM8K dataset anddeveloped the MR-GSM8K benchmark. Our extensive analysis includes severalstate-of-the-art models from both open-source and commercial domains, uncovering fundamental deficiencies in their training and evaluation methodologies.Specifically, we found the OpenAI o1 models which possess characteristics of"system-2" thinking excel the other SOTA models by more than 20 absolute pointsin our benchmark, supporting our deficiency hypothesis.

308. CEB: Compositional Evaluation Benchmark for Fairness in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30167 abstract： As Large Language Models (LLMs) are increasingly deployed to handle various natural language processing (NLP) tasks, concerns regarding the potential negative societal impacts of LLM-generated content have also arisen. To evaluate the biases exhibited by LLMs, researchers have recently proposed a variety of datasets. However, existing bias evaluation efforts often focus on only a particular type of bias and employ inconsistent evaluation metrics, leading to difficulties in comparison across different datasets and LLMs. To address these limitations, we collect a variety of datasets designed for the bias evaluation of LLMs, and further propose CEB, a Compositional Evaluation Bechmark that covers different types of bias across different social groups and tasks. The curation of CEB is based on our newly proposed compositional taxonomy, which characterizes each dataset from three dimensions: bias types, social groups, and tasks. By combining the three dimensions, we develop a comprehensive evaluation strategy for the bias in LLMs. Our experiments demonstrate that the levels of bias vary across these dimensions, thereby providing guidance for the development of specific bias mitigation methods.

309. Explanations of GNN on Evolving Graphs via Axiomatic Layer edges

链接：https://iclr.cc/virtual/2025/poster/28293 abstract： Graphs are ubiquitous in social networks, chemical molecules, and financial data, where Graph Neural Networks (GNNs) achieve superior predictive accuracy. Graphs can beevolving, while understanding how GNN predictions respond to the evolution provides significant insight and trust. We explore the problem of explaining evolving GNN predictions due to continuously changing edge weights.We introduce a layer edge-based explanation to balanceexplanation fidelity and interpretability.We propose a novel framework to address the challenges of axiomatic attribution and the entanglement of multiple computational graph paths due to continuous change of edge weights. We first design an axiomatic attribution of the evolution of the model prediction to message flows, then develop Shapley value to fairly map message flow contributions to layer edges.We formulate a novel optimization problem to find the critical layer edges based on KL-divergence minimization. Extensive experiments on eight datasets for node classification, link prediction, and graph classification tasks with evolving graphs demonstrate the better fidelity and interpretability of the proposed method over the baseline methods. The code is available at https://github.com/yazhengliu/Axiomatic-Layer-Edges/tree/main.

310. ARB-LLM: Alternating Refined Binarizations for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29217 abstract： Large Language Models (LLMs) have greatly pushed forward advancements in natural language processing, yet their high memory and computational demands hinder practical deployment. Binarization, as an effective compression technique, can shrink model weights to just 1 bit, significantly reducing the high demands on computation and memory. However, current binarization methods struggle to narrow the distribution gap between binarized and full-precision weights, while also overlooking the column deviation in LLM weight distribution. To tackle these issues, we propose ARB-LLM, a novel 1-bit post-training quantization (PTQ) technique tailored for LLMs. To narrow the distribution shift between binarized and full-precision weights, we first design an alternating refined binarization (ARB) algorithm to progressively update the binarization parameters, which significantly reduces the quantization error. Moreover, considering the pivot role of calibration data and the column deviation in LLM weights, we further extend ARB to ARB-X and ARB-RC. In addition, we refine the weight partition strategy with column-group bitmap (CGB), which further enhance performance. Equipping ARB-X and ARB-RC with CGB, we obtain ARB-LLM$*{\text{X}}$ and ARB-LLM$* {\text{RC}} $ respectively, which significantly outperform state-of-the-art (SOTA) binarization methods for LLMs.As a binary PTQ method, our ARB-LLM$ _{\text{RC}} $ is the first to surpass FP16 models of the same size. Code: https://github.com/ZHITENGLI/ARB-LLM.

311. Unlocking Efficient, Scalable, and Continual Knowledge Editing with Basis-Level Representation Fine-Tuning

链接：https://iclr.cc/virtual/2025/poster/29762 abstract： Large language models (LLMs) have achieved remarkable performance on vari-ous natural language tasks. However, they are trained on static corpora and theirknowledge can become outdated quickly in the fast-changing world. This moti-vates the development of knowledge editing methods designed to update certainknowledge in LLMs without changing unrelated others. To make selective edits,previous efforts often sought to update a small amount of parameters in some spe-cific layer(s) of a LLM. Nonetheless, in challenging scenarios, they still fall shortin making successful edits while preserving knowledge irrelevant to the updatessimultaneously, resulting in a notable editing-locality trade-off. In this work, wequestion if the trade-offs are caused by the fact that parameter-based updates havea global effect, i.e., edited parameters affect all inputs indiscriminately. In light ofthis, we explore the feasibility of representation fine-tuning, which applied somelinear update to a few representations in a learned subspace, for knowledge edit-ing. While being effective to enhance an LLM’s general ability as demonstrated inthe previous work, we theoretically show that this linear update imposes a tensionin editing-locality trade-off. Subsequently, BaFT is proposed to break the linear-ity. BaFT computes a weight

for each basis that spans a dimension of the subspacebased on the input representation. This input-dependent weighting mechanism al-lows BaFT to manage different types of knowledge in an adaptive way, therebyachieving a better editing-locality trade-off. Experiments on three LLMs with fiveediting benchmarks in diverse scenarios show the superiority of our method.

312. Advantage-Guided Distillation for Preference Alignment in Small Language Models

链接：https://iclr.cc/virtual/2025/poster/27759 abstract： Alignment techniques enable Large Language Models (LLMs) to generate outputs that align with human preferences and play a crucial role in their effectiveness. However, their impact often diminishes when applied to Small Language Models (SLMs), likely due to the limited capacity of these models. Instead of directly applying existing alignment techniques to SLMs, we propose to utilize a well-aligned teacher LLM to guide the alignment process for these models, thereby facilitating the transfer of the teacher's knowledge of human preferences to the student model. To achieve this, we first explore a straightforward approach, Dual-Constrained Knowledge Distillation (DCKD), that employs knowledge distillation with two KL-divergence constraints from the aligned teacher to the unaligned student. To further enhance the student's ability to distinguish between preferred and dispreferred responses, we then propose Advantage-Guided Distillation for Preference Alignment (ADPA), which leverages an advantage function from the aligned teacher to deliver more nuanced, distribution-level reward signals for the student's alignment. Our experimental results show that these two approaches appreciably improve the alignment of SLMs and narrow the performance gap with larger counterparts. Among them, ADPA demonstrates superior performance and achieves even greater effectiveness when integrated with DCKD. Our code is available at https://github.com/SLIT-AI/ADPA .

313. Dynamic Modeling of Patients, Modalities and Tasks via Multi-modal Multi-task Mixture of Experts

链接：https://iclr.cc/virtual/2025/poster/29884 abstract： Multi-modal multi-task learning holds significant promise in tackling complex diagnostic tasks and many significant medical imaging problems. It fulfills the needs in real-world diagnosis protocol to leverage information from different data sources and simultaneously perform mutually informative tasks. However, medical imaging domains introduce two key challenges: dynamic modality fusion and modality-task dependence. The quality and amount of task-related information from different modalities could vary significantly across patient samples, due to biological and demographic factors. Traditional fusion methods apply fixed combination strategies that fail to capture this dynamic relationship, potentially underutilizing modalities that carry stronger diagnostic signals for specific patients. Additionally, different clinical tasks may require dynamic feature selection and combination from various modalities, a phenomenon we term “modality-task dependence.” To address these issues, we propose M4oE, a novel Multi-modal Multi-task Mixture of Experts framework for precise Medical diagnosis. M4oE comprises Modality-Specific (MSoE) modules and a Modality-shared Modality-Task MoE (MToE) module. With collaboration from both modules, our model dynamically decomposes and learns distinct and shared information from different modalities and achieves dynamic fusion. MToE provides a joint probability model of modalities and tasks by using experts as a link and encourages experts to learn modality-task dependence via conditional mutual information loss. By doing so, M4oE offers sample and population-level interpretability of modality contributions. We evaluate M4oE on four public multi-modal medical benchmark datasets for solving two important medical diagnostic problems including breast cancer screening and retinal disease diagnosis. Results demonstrate our method's superiority over state-of-the-art methods under different metrics of classification and segmentation tasks like Accuracy, AUROC, AUPRC, and DICE.

314. Shot2Story: A New Benchmark for Comprehensive Understanding of Multi-shot Videos

链接：https://iclr.cc/virtual/2025/poster/30335 abstract： A short clip of video may contain progression of multiple events and an interesting story line. A human need to capture both the event in every shot and associate them together to understand the story behind it. In this work, we present a new multi-shot video understanding benchmark \dataset with detailed shot-level captions, comprehensive video summaries and question-answering pairs. To facilitate better semantic understanding of videos, we provide captions for both visual signals and human narrations. We design several distinct tasks including single-shot video captioning, multi-shot video summarization, and multi-shot video question answering. Preliminary experiments show some challenges to generate a long and comprehensive video summary for multi-shot videos. Nevertheless, the generated imperfect summaries can already achieve competitive performance on existing video understanding tasks such as video question-answering, promoting an under-explored setting of video understanding with detailed summaries.

315. Limits of Deep Learning: Sequence Modeling through the Lens of Complexity Theory

链接：https://iclr.cc/virtual/2025/poster/30446 abstract： Despite their successes, deep learning models struggle with tasks requiring complex reasoning and function composition. We present a theoretical and empirical investigation into the limitations of Structured State Space Models (SSMs) and Transformers in such tasks. We prove that one-layer SSMs cannot efficiently perform function composition over large domains without impractically large state sizes, and even with Chain-of-Thought prompting, they require a number of steps that scale unfavorably with the complexity of the function composition. Also, the language of a finite-precision SSM is within the class of regular languages. Our experiments corroborate these theoretical

findings. Evaluating models on tasks including various function composition settings, multi-digit multiplication, dynamic programming, and Einstein's puzzle, we find significant performance degradation even with advanced prompting techniques. Models often resort to shortcuts, leading to compounding errors. These findings highlight fundamental barriers within current deep learning architectures rooted in their computational capacities. We underscore the need for innovative solutions to transcend these constraints and achieve reliable multi-step reasoning and compositional task-solving, which is critical for advancing toward general artificial intelligence.

316. Fragment and Geometry Aware Tokenization of Molecules for Structure-Based Drug Design Using Language Models

链接：https://iclr.cc/virtual/2025/poster/28470 abstract： Structure-based drug design (SBDD) is crucial for developing specific and effective therapeutics against protein targets but remains challenging due to complex protein-ligand interactions and vast chemical space. Although language models (LMs) have excelled in natural language processing, their application in SBDD is underexplored. To bridge this gap, we introduce a method, known as Frag2Seq, to apply LMs to SBDD by generating molecules in a fragment-based manner in which fragments correspond to functional modules. We transform 3D molecules into fragment-informed sequences using $SE(3)$-equivariant molecule and fragment local frames, extracting $SE(3)$-invariant sequences that preserve geometric information of 3D fragments. Furthermore, we incorporate protein pocket embeddings obtained from a pre-trained inverse folding model into the LMs via cross-attention to capture protein-ligand interaction, enabling effective target-aware molecule generation. Benefiting from employing LMs with fragment-based generation and effective protein context encoding, our model achieves the best performance on binding vina score and chemical properties such as QED and Lipinski, which shows our model’s efficacy in generating drug-like ligands with higher binding affinity against target proteins. Moreover, our method also exhibits higher sampling efficiency compared to atom-based autoregressive and diffusion baselines with at most $\times 300$ speedup. The code will be made publicly available at https://github.com/divelab/AIRS/tree/main/OpenMI/Frag2Seq.

317. ASTrA: Adversarial Self-supervised Training with Adaptive-Attacks

链接：https://iclr.cc/virtual/2025/poster/29209 abstract： Existing self-supervised adversarial training (self-AT) methods rely on hand-crafted adversarial attack strategies for PGD attacks, which fail to adapt to the evolving learning dynamics of the model and do not account for instance-specific characteristics of images. This results in sub-optimal adversarial robustness and limits the alignment between clean and adversarial data distributions. To address this, we propose $\textit{ASTrA}$ ($\textbf{A}$dversarial $\textbf{S}$elf-supervised $\textbf{Tr}$aining with $\textbf{A}$daptive-Attacks), a novel framework introducing a learnable, self-supervised attack strategy network that autonomously discovers optimal attack parameters through exploration-exploitation in a single training episode. ASTrA leverages a reward mechanism based on contrastive loss, optimized with REINFORCE, enabling adaptive attack strategies without labeled data or additional hyperparameters. We further introduce a mixed contrastive objective to align the distribution of clean and adversarial examples in representation space. ASTrA achieves state-of-the-art results on CIFAR10, CIFAR100, and STL10 while integrating seamlessly as a plug-and-play module for other self-AT methods. ASTrA shows scalability to larger datasets, demonstrates strong semi-supervised performance, and is resilient to robust overfitting, backed by explainability analysis on optimal attack strategies. Project page for source code and other details at https://prakashchhipa.github.io/projects/ASTrA.

318. MambaQuant: Quantizing the Mamba Family with Variance Aligned Rotation Methods

链接：https://iclr.cc/virtual/2025/poster/30065 abstract： Mamba is an efficient sequence model that rivals Transformers and demonstrates significant potential as a foundational architecture for various tasks. Quantization is commonly used in neural networks to reduce model size and computational latency. However, applying quantization to Mamba remains underexplored, and existing quantization methods, which have been effective for CNN and Transformer models, appear inadequate for Mamba models (e.g., Quarot suffers a 21% accuracy drop on Vim-T$\dagger$ even under W8A8). We have pioneered the exploration of this issue and identified several key challenges. First, significant outliers arepresent in gate projections, output projections, and matrix multiplications. Second, Mamba’s unique parallel scan further amplifies these outliers, leading to uneven and heavy-tailed data distributions. Third, even with the application of the Hadamard transform, the variance across channels in weights and activations still remains inconsistent. To these ends, we propose MambaQuant, a post-training quantization (PTQ) framework consisting of: 1) Karhunen-Lo`eve Transformation (KLT) enhanced rotation, rendering the rotation matrix adaptable to diverse channel distributions. 2) Smooth-Fused rotation, which equalizes channel variances and can merge additional parameters into model weights. Experiments show that MambaQuant can quantize both weights and activations into 8-bit with less than 1% accuracy loss for Mamba-based vision and language tasks. To our knowledge, MambaQuant is the first comprehensive PTQ design for the Mamba family, paving the way for further advancements in its application.

319. Cross-Attention Head Position Patterns Can Align with Human Visual Concepts in Text-to-Image Generative Models

链接：https://iclr.cc/virtual/2025/poster/31175 abstract： Recent text-to-image diffusion models leverage cross-attention layers, which have been effectively utilized to enhance a range of visual generative tasks. However, our understanding of cross-attention

layers remains somewhat limited. In this study, we introduce a mechanistic interpretability approach for diffusion models by constructing Head Relevance Vectors (HRVs) that align with human-specified visual concepts. An HRV for a given visual concept has a length equal to the total number of cross-attention heads, with each element representing the importance of the corresponding head for the given visual concept. To validate HRVs as interpretable features, we develop an ordered weakening analysis that demonstrates their effectiveness. Furthermore, we propose concept strengthening and concept adjusting methods and apply them to enhance three visual generative tasks. Our results show that HRVs can reduce misinterpretations of polysemous words in image generation, successfully modify five challenging attributes in image editing, and mitigate catastrophic neglect in multi-concept generation. Overall, our work provides an advancement in understanding cross-attention layers and introduces new approaches for fine-controlling these layers at the head level.

320. OSTQuant: Refining Large Language Model Quantization with Orthogonal and Scaling Transformations for Better Distribution Fitting

链接：https://iclr.cc/virtual/2025/poster/28216 abstract： Post-training quantization (PTQ) has emerged as a widely adopted technique for compressing and accelerating Large Language Models (LLMs).The major challenge in LLM quantization is that uneven and heavy-tailed data distributions can expand the quantization range, thereby reducing bit precision for most values.Recent methods attempt to eliminate outliers and balance inter-channel differences by employing linear transformations; however, they remain heuristic and are often overlook optimizing the data distribution across the entire quantization space.In this paper, we introduce Quantization Space Utilization Rate (QSUR), a novel metric that effectively assesses the quantizability of transformed data by measuring the space utilization of the data in the quantization space. We complement QSUR with mathematical derivations that examine the effects and limitations of various transformations, guiding our development of Orthogonal and Scaling Transformation-based Quantization (OSTQuant). OSTQuant employs a learnable equivalent transformation, consisting of an orthogonal transformation and a scaling transformation, to optimize the distributions of weights and activations across the entire quantization space. Futhermore, we propose the KL-Top loss function, designed to mitigate noise during optimization while retaining richer semantic information within the limited calibration data imposed by PTQ.OSTQuant outperforms existing work on various LLMs and benchmarks. In the W4-only setting, it retains 99.5\% of the floating-point accuracy. In the more challenging W4A4KV4 configuration, OSTQuant reduces the performance gap by 32\% on the LLaMA-3-8B model compared to state-of-the-art methods. Code will be available.

321. GOttack: Universal Adversarial Attacks on Graph Neural Networks via Graph Orbits Learning

链接：https://iclr.cc/virtual/2025/poster/29256 abstract： Graph Neural Networks (GNNs) have demonstrated superior performance in node classification tasks across diverse applications. However, their vulnerability to adversarial attacks, where minor perturbations can mislead model predictions, poses significant challenges. This study introduces GOttack, a novel adversarial attack framework that exploits the topological structure of graphs to undermine the integrity of GNN predictions systematically. By defining a topology-aware method to manipulate graph orbits, our approach generates adversarial modifications that are both subtle and effective, posing a severe test to the robustness of GNNs. We evaluate the efficacy of GOttack across multiple prominent GNN architectures using standard benchmark datasets. Our results show that GOttack outperforms existing state-of-the-art adversarial techniques and completes training in approximately 55% of the time required by the fastest competing model, achieving the highest average misclassification rate in 155 tasks. This work not only sheds light on the susceptibility of GNNs to structured adversarial attacks but also shows that certain topological patterns may play a significant role in the underlying robustness of the GNNs. Our Python implementation is shared at https://github.com/cakcora/GOttack.

322. Misspecified Q-Learning with Sparse Linear Function Approximation: Tight Bounds on Approximation Error

链接：https://iclr.cc/virtual/2025/poster/28411 abstract： The recent work by Dong and Yang (2023) showed for misspecified sparse linear bandits, one can obtain an $O(\epsilon)$-optimal policy using a polynomial number of samples when the sparsity is a constant, where ϵ is the misspecification error. This result is in sharp contrast to misspecified linear bandits without sparsity, which require an exponential number of samples to get the same guarantee. In order to study whether the analog result is possible in the reinforcement learning setting, we consider the following problem: assuming the optimal Q-function is a d-dimensional linear function with sparsity k and misspecification error ϵ, whether we can obtain an $O(\epsilon)$-optimal policy using number of samples polynomially in the feature dimension d. We first demonstrate why the standard approach based on Bellman backup or the existing optimistic value function elimination approach such as OLIVE (Jiang et al., 2017) achieves suboptimal guarantees for this problem. We then design a novel elimination-based algorithm to show one can obtain an $O(H\epsilon)$-optimal policy with sample complexity polynomially in the feature dimension d and planning horizon H. Lastly, we complement our upper bound with an $\tilde \Omega(H\epsilon)$ suboptimality lower bound, giving a complete picture of this problem.

323. MrSteve: Instruction-Following Agents in Minecraft with What-Where-When Memory

链接：https://iclr.cc/virtual/2025/poster/30503 abstract： Significant advances have been made in developing general-purpose

embodied AI in environments like Minecraft through the adoption of LLM-augmented hierarchical approaches. While these approaches, which combine high-level planners with low-level controllers, show promise, low-level controllers frequently become performance bottlenecks due to repeated failures. In this paper, we argue that the primary cause of failure in many low-level controllers is the absence of an episodic memory system. To address this, we introduce MrSteve (Memory Recall Steve), a novel low-level controller equipped with Place Event Memory (PEM), a form of episodic memory that captures what, where, and when information from episodes. This directly addresses the main limitation of the popular low-level controller, Steve-1. Unlike previous models that rely on short-term memory, PEM organizes spatial and event-based data, enabling efficient recall and navigation in long-horizon tasks. Additionally, we propose an Exploration Strategy and a Memory-Augmented Task Solving Framework, allowing agents to alternate between exploration and task-solving based on recalled events. Our approach significantly improves task-solving and exploration efficiency compared to existing methods. We will release our code and demos on the project page: https://sites.google.com/view/mr-steve.

324. Regret-Optimal List Replicable Bandit Learning: Matching Upper and Lower Bounds

链接：https://iclr.cc/virtual/2025/poster/31254 abstract： This paper investigates *list replicability* [Dixon et al., 2023] in the context of multi-armed (also linear) bandits (MAB). We define an algorithm A for MAB to be (ℓ,δ)-list replicable if with probability at least $1-\delta$, A has at most ℓ traces in independent executions even with different random bits, where a trace means sequence of arms played during an execution. For k-armed bandits, although the total number of traces can be $\Omega(k^T)$ for a time horizon T, we present several surprising upper bounds that either independent of or logarithmic of T: (1) a $(2^{k},\delta)$-list replicable algorithm with near-optimal regret, $\widetilde{O}({\sqrt{kT}})$, (2) a $(O(k/\delta),\delta)$-list replicable algorithm with regret $\widetilde{O}\left(\frac{k}{\delta}\sqrt{kT}\right)$, (3) a $((k+1)^{B-1}, \delta)$-list replicable algorithm with regret $\widetilde{O}(k^{\frac{3}{2}}T^{{\frac{1}{2}}+2^{-(B+1)}})$ for any integer $B>1$. On the other hand, for the sublinear regret regime, we establish a matching lower bound on the list complexity (parameter ℓ). We prove that there is no $(k-1,\delta)$-list replicable algorithm with $o(T)$-regret. This is optimal in list complexity in the sub-linear regret regime as there is a $(k, 0)$-list replicable algorithm with $O(T^{2/3})$-regret. We further show that for linear bandits with d-dimensional features, there is a $\widetilde{O}(d^2T^{1/2+2^{-(B+1)}})$-regret algorithm with $((2d+1)^{B-1},\delta)$-list replicability, for $B>1$, even when the number of possible arms can be infinite.

325. Efficient Imitation under Misspecification

链接：https://iclr.cc/virtual/2025/poster/28859 abstract： We consider the problem of imitation learning under misspecification: settings where the learner is fundamentally unable to replicate expert behavior everywhere. This is often true in practice due to differences in observation space and action space expressiveness (e.g. perceptual or morphological differences between robots and humans). Given the learner must make some mistakes in the misspecified setting, interaction with the environment is fundamentally required to figure out which mistakes are particularly costly and lead to compounding errors. However, given the computational cost and safety concerns inherent in interaction, we'd like to perform as little of it as possible while ensuring we've learned a strong policy. Accordingly, prior work has proposed a flavor of efficient inverse reinforcement learning algorithms that merely perform a computationally efficient local search procedure with strong guarantees in the realizable setting. We first prove that under a novel structural condition we term reward-agnostic policy completeness, these sorts of local-search based IRL algorithms are able to avoid compounding errors. We then consider the question of where we should perform local search in the first place, given the learner may not be able to "walk on a tightrope" as well as the expert in the misspecified setting. We prove that in the misspecified setting, it is beneficial to broaden the set of states on which local search is performed to include those reachable by good policies the learner can actually play. We then experimentally explore a variety of sources of misspecification and how offline data can be used to effectively broaden where we perform local search from.

326. Enhancing Robust Fairness via Confusional Spectral Regularization

链接：https://iclr.cc/virtual/2025/poster/28515 abstract： Recent research has highlighted a critical issue known as ``robust fairness", where robust accuracy varies significantly across different classes, undermining the reliability of deep neural networks (DNNs). A common approach to address this has been to dynamically reweight classes during training, giving more weight to those with lower empirical robust performance. However, we find there is a divergence of class-wise robust performance between training set and testing set, which limits the effectiveness of these explicit reweighting methods, indicating the need for a principled alternative.In this work, we derive a robust generalization bound for the worst-class robust error within the PAC-Bayesian framework, accounting for unknown data distributions. Our analysis shows that the worst-class robust error is influenced by two main factors: the spectral norm of the empirical robust confusion matrix and the information embedded in the model and training set. While the latter has been extensively studied, we propose a novel regularization technique targeting the spectral norm of the robust confusion matrix to improve worst-class robust accuracy and enhance robust fairness.We validate our approach through comprehensive experiments on various datasets and models, demonstrating its effectiveness in enhancing robust fairness.

327. Mini-Monkey: Alleviating the Semantic Sawtooth Effect for Lightweight MLLMs via Complementary Image Pyramid

链接：https://iclr.cc/virtual/2025/poster/30854 abstract： Recently, scaling images to high resolution has received much

attention in multimodal large language models (MLLMs). Most existing practices adopt a sliding-window-style cropping strategy to adapt to resolution increase. Such a cropping strategy, however, can easily cut off objects and connected regions, which introduces semantic discontinuity and therefore impedes MLLMs from recognizing small or irregularly shaped objects or text, leading to a phenomenon we call the semantic sawtooth effect. This effect is particularly evident in lightweight MLLMs. To address this issue, we introduce a Complementary Image Pyramid (CIP), a simple, effective, and plug-and-play solution designed to mitigate semantic discontinuity during high-resolution image processing. In particular, CIP dynamically constructs an image pyramid to provide complementary semantic information for the cropping-based MLLMs, enabling it rich acquire semantics at all levels. Furthermore, we introduce a Scale Compression Mechanism (SCM) to reduce the additional computational overhead by compressing the redundant visual tokens. Our experiments demonstrate that CIP can consistently enhance the performance across diverse architectures (e.g., MiniCPM-V-2, InternVL2, and LLaVA-OneVision), various model capacity (1B$\rightarrow$8B), and different usage configurations (training-free and fine-tuning). Leveraging the proposed CIP and SCM, we introduce a lightweight MLLM, Mini-Monkey, which achieves remarkable performance in both general multimodal understanding and document understanding. On the OCRBench, the 2B-version Mini-Monkey even surpasses the 8B model InternVL2-8B by 12 score. Additionally, training Mini-Monkey is cheap, requiring only eight RTX 3090 GPUs. Code and models are available at https://github.com/Yuliang-Liu/Monkey.

328. ThermalGaussian: Thermal 3D Gaussian Splatting

链接：https://iclr.cc/virtual/2025/poster/27712 abstract： Thermography is especially valuable for the military and other users of surveillance cameras. Some recent methods based on Neural Radiance Fields (NeRF) are proposed to reconstruct the thermal scenes in 3D from a set of thermal and RGB images. However, unlike NeRF, 3D Gaussian splatting (3DGS) prevails due to its rapid training and real-time rendering. In this work, we propose ThermalGaussian, the first thermal 3DGS approach capable of rendering high-quality images in RGB and thermal modalities. We first calibrate the RGB camera and the thermal camera to ensure that both modalities are accurately aligned. Subsequently, we use the registered images to learn the multimodal 3D Gaussians. To prevent the overfitting of any single modality, we introduce several multimodal regularization constraints. We also develop smoothing constraints tailored to the physical characteristics of the thermal modality.Besides, we contribute a real-world dataset named RGBT-Scenes, captured by a hand-hold thermal-infrared camera, facilitating future research on thermal scene reconstruction. We conduct comprehensive experiments to show that ThermalGaussian achieves photorealistic rendering of thermal images and improves the rendering quality of RGB images. With the proposed multimodal regularization constraints, we also reduced the model's storage cost by 90\%. Our project page is at https://thermalgaussian.github.io/.

329. FreeCG: Free the Design Space of Clebsch-Gordan Transform for Machine Learning Force Fields

链接：https://iclr.cc/virtual/2025/poster/28107 abstract： Machine Learning Force Fields (MLFFs) are of great importance for chemistry, physics, materials science, and many other related fields. The Clebsch–Gordan transform (CG transform) effectively encodes many-body interactions and is thus an important building block for many models of MLFFs. However, the permutation-equivariance requirement of MLFFs limits the design space of CG transform, that is, intensive CG transform has to be conducted for each neighboring edge and the operations should be performed in the same manner for all edges. Freeing up the design space can greatly improve the model's expressiveness while simultaneously decreasing computational demands. To reach this goal, we utilize a mathematical proposition, invariance transitivity, to show that implementing the CG transform layer on the permutation-invariant abstract edges allows complete freedom in the design of the layer without compromising the overall permutation equivariance. Developing on this free design space, we further propose group CG transform with sparse path, abstract edges shuffling, and attention enhancer to form a powerful and efficient CG transform layer. Our method, known as FreeCG, achieves state-of-the-art (SOTA) results in force prediction for MD17, rMD17, MD22, and is well extended to property prediction in QM9 datasets with several improvements greater than 15% and the maximum beyond 20%. The extensive real-world applications showcase high practicality. FreeCG introduces a novel paradigm for carrying out efficient and expressive CG transform in future geometric network designs. To demonstrate this, the recent SOTA, QuinNet, is also enhanced under our paradigm. Code: https://github.com/ShihaoShao-GH/FreeCG.

330. From Layers to States: A State Space Model Perspective to Deep Neural Network Layer Dynamics

链接：https://iclr.cc/virtual/2025/poster/28443 abstract： The depth of neural networks is a critical factor for their capability, with deeper models often demonstrating superior performance. Motivated by this, significant efforts have been made to enhance layer aggregation - reusing information from previous layers to better extract features at the current layer, to improve the representational power of deep neural networks. However, previous works have primarily addressed this problem from a discrete-state perspective which is not suitable as the number of network layers grows. This paper novelly treats the outputs from layers as states of a continuous process and considers leveraging the state space model (SSM) to design the aggregation of layers in very deep neural networks. Moreover, inspired by its advancements in modeling long sequences, the Selective State Space Models (S6) is employed to design a new module called Selective State Space Model Layer Aggregation (S6LA). This module aims to combine traditional CNN or transformer architectures within a sequential framework, enhancing the representational capabilities of state-of-the-art vision networks. Extensive experiments show that S6LA delivers substantial improvements in both image classification and detection tasks, highlighting the potential of integrating SSMs with contemporary deep learning techniques.

331. From Risk to Uncertainty: Generating Predictive Uncertainty Measures via Bayesian Estimation

链接：https://iclr.cc/virtual/2025/poster/29041 abstract： There are various measures of predictive uncertainty in the literature, but their relationships to each other remain unclear. This paper uses a decomposition of statistical pointwise risk into components associated with different sources of predictive uncertainty: namely, aleatoric uncertainty (inherent data variability) and epistemic uncertainty (model-related uncertainty). Together with Bayesian methods applied as approximations, we build a framework that allows one to generate different predictive uncertainty measures.We validate measures, derived from our framework on image datasets by evaluating its performance in detecting out-of-distribution and misclassified instances using the AUROC metric. The experimental results confirm that the measures derived from our framework are useful for the considered downstream tasks.

332. Enhancing the Scalability and Applicability of Kohn-Sham Hamiltonians for Molecular Systems

链接：https://iclr.cc/virtual/2025/poster/28011 abstract： Density Functional Theory (DFT) is a pivotal method within quantum chemistry and materials science, with its core involving the construction and solution of the Kohn-Sham Hamiltonian. Despite its importance, the application of DFT is frequently limited by the substantial computational resources required to construct the Kohn-Sham Hamiltonian. In response to these limitations, current research has employed deep-learning models to efficiently predict molecular and solid Hamiltonians, with roto-translational symmetries encoded in their neural networks. However, the scalability of prior models may be problematic when applied to large molecules, resulting in non-physical predictions of ground-state properties. In this study, we generate a substantially larger training set (PubChemQH) than used previously and use it to create a scalable model for DFT calculations with physical accuracy. For our model, we introduce a loss function derived from physical principles, which we call Wavefunction Alignment Loss (WALoss). WALoss involves performing a basis change on the predicted Hamiltonian to align it with the observed one; thus, the resulting differences can serve as a surrogate for orbital energy differences, allowing models to make better predictions for molecular orbitals and total energies than previously possible. WALoss also substantially accelerates self-consistent-field (SCF) DFT calculations. Here, we show it achieves a reduction in total energy prediction error by a factor of 1347 and an SCF calculation speed-up by a factor of 18\%. These substantial improvements set new benchmarks for achieving accurate and applicable predictions in larger molecular systems.

333. Computationally Efficient RL under Linear Bellman Completeness for Deterministic Dynamics

链接：https://iclr.cc/virtual/2025/poster/28728 abstract： We study computationally and statistically efficient Reinforcement Learning algorithms for the linear Bellman Complete setting. This setting uses linear function approximation to capture value functions and unifies existing models like linear Markov Decision Processes (MDP) and Linear Quadratic Regulators (LQR). While it is known from the prior works that this setting is statistically tractable, it remained open whether a computationally efficient algorithm exists. Our work provides a computationally efficient algorithm for the linear Bellman complete setting that works for MDPs with large action spaces, random initial states, and random rewards but relies on the underlying dynamics to be deterministic. Our approach is based on randomization: we inject random noise into least squares regression problems to perform optimistic value iteration. Our key technical contribution is to carefully design the noise to only act in the null space of the training data to ensure optimism while circumventing a subtle error amplification issue.

334. Optimization with Access to Auxiliary Information

链接：https://iclr.cc/virtual/2025/poster/31502 abstract： We investigate the fundamental optimization question of minimizing a \emph{target} function $f(x)$, whose gradients are expensive to compute or have limited availability, given access to some \emph{auxiliary} side function $h(x)$ whose gradients are cheap or more available. This formulation captures many settings of practical relevance, such as i) re-using batches in SGD, ii) transfer learning, iii) federated learning, iv) training with compressed models/dropout, etcetera. We propose two generic new algorithms that apply in all these settings; we also prove that we can benefit from this framework under the Hessian similarity assumption between the target and side information. A benefit is obtained when this similarity measure is small; we also show a potential benefit from stochasticity when the auxiliary noise is correlated with that of the target function.

335. Bandit Learning in Matching Markets with Indifference

链接：https://iclr.cc/virtual/2025/poster/30837 abstract： A rich line of recent works studies how participants in matching markets learn their unknown preferences through iterative interactions with each other. The two sides of participants in the market can be respectively formulated as players and arms in the bandit problem. To ensure market stability, the objective is to minimize the stable regret of each player. Though existing works provide significant theoretical upper bounds for players' stable regret, the results heavily rely on the assumption that each participant has a strict preference ranking. However, in real applications, multiple candidates (e.g., workers in the labor market and students in school admission) usually demonstrate comparable performance levels, making it challenging for participants (e.g., employers and schools) to differentiate and rank their preferences. To deal with the potential indifferent preferences, we propose an adaptive exploration algorithm based on

arm-guided Gale-Shapley (AE-AGS). We show that its stable regret is of order $O(NK \log T / \Delta^2)$, where N is the number of players, K the number of arms, T the total time horizon, and Δ the minimum non-zero preference gap. Extensive experiments demonstrate the algorithm's effectiveness in handling such complex situations and its consistent superiority over baselines.

336. Grounding Video Models to Actions through Goal Conditioned Exploration

链接：https://iclr.cc/virtual/2025/poster/30300 abstract： Large video models, pretrained on massive quantities of amount of Internet video, provide a rich source of physical knowledge about the dynamics and motions of objects and tasks.However, video models are not grounded in the embodiment of an agent, and do not describe how to actuate the world to reach the visual states depicted in a video.To tackle this problem, current methods use a separate vision-based inverse dynamic model trained on embodiment-specific data to map image states to actions. Gathering data to train such a model is often expensive and challenging, and this model is limited to visual settings similar to the ones in which data is available.In this paper, we investigate how to directly ground video models to continuous actions through self-exploration in the embodied environment -- using generated video states as visual goals for exploration.We propose a framework that uses trajectory level action generation in combination with video guidance toenable an agent to solve complex tasks without any external supervision, e.g., rewards, action labels, or segmentation masks.We validate the proposed approach on 8 tasks in Libero, 6 tasks in MetaWorld, 4 tasks in Calvin, and 12 tasks in iThor Visual Navigation. We show how our approach is on par with or even surpasses multiple behavior cloning baselines trained on expert demonstrations while without requiring any action annotations.

337. Proteina: Scaling Flow-based Protein Structure Generative Models

链接：https://iclr.cc/virtual/2025/poster/29538 abstract： Recently, diffusion- and flow-based generative models of protein structures have emerged as a powerful tool for de novo protein design. Here, we develop *Proteina*, a new large-scale flow-based protein backbone generator that utilizes hierarchical fold class labels for conditioning and relies on a tailored scalable transformer architecture with up to $5\times$ as many parameters as previous models. To meaningfully quantify performance, we introduce a new set of metrics that directly measure the distributional similarity of generated proteins with reference sets, complementing existing metrics. We further explore scaling training data to millions of synthetic protein structures and explore improved training and sampling recipes adapted to protein backbone generation. This includes fine-tuning strategies like LoRA for protein backbones, new guidance methods like classifier-free guidance and autoguidance for protein backbones, and new adjusted training objectives. Proteina achieves state-of-the-art performance on de novo protein backbone design and produces diverse and designable proteins at unprecedented length, up to 800 residues. The hierarchical conditioning offers novel control, enabling high-level secondary-structure guidance as well as low-level fold-specific generation.

338. On Speeding Up Language Model Evaluation

链接：https://iclr.cc/virtual/2025/poster/31065 abstract： Developing prompt-based methods with Large Language Models (LLMs) requires making numerous decisions, which give rise to a combinatorial search problem over hyper-parameters. This exhaustive evaluation can be time-consuming and costly. In this paper, we propose an \textit{adaptive} approach to explore this space. We are exploiting the fact that often only few samples are needed to identify clearly superior or inferior settings, and that many evaluation tests are highly correlated. We lean on multi-armed bandits to sequentially identify the next (method, validation sample)-pair to evaluate and utilize low-rank matrix factorization to fill in missing evaluations. We carefully assess the efficacy of our approach on several competitive benchmark problems and show that it can identify the top-performing method using only 5-15% of the typical resources---resulting in 85-95% LLM cost savings. Our code is available at https://github.com/kilian-group/banditeval.

339. You Only Sample Once: Taming One-Step Text-to-Image Synthesis by Self-Cooperative Diffusion GANs

链接：https://iclr.cc/virtual/2025/poster/29561 abstract： Recently, some works have tried to combine diffusion and Generative Adversarial Networks (GANs) to alleviate the computational cost of the iterative denoising inference in Diffusion Models (DMs). However, existing works in this line suffer from either training instability and mode collapse or subpar one-step generation learning efficiency. To address these issues, we introduce YOSO, a novel generative model designed for rapid, scalable, and high-fidelity one-step image synthesis with high training stability and mode coverage. Specifically, we smooth the adversarial divergence by the denoising generator itself, performing self-cooperative learning. We show that our method can serve as a one-step generation model training from scratch with competitive performance. Moreover, we extend our YOSO to one-step text-to-image generation based on pre-trained models by several effective training techniques (i.e., latent perceptual loss and latent discriminator for efficient training along with the latent DMs; the informative prior initialization (IPI), and the quick adaption stage for fixing the flawed noise scheduler). Experimental results show that YOSO achieves the state-of-the-art one-step generation performance even with Low-Rank Adaptation (LoRA) fine-tuning.In particular, we show that the YOSO-PixArt-α can generate images in one step trained on 512 resolution, with the capability of adapting to 1024 resolution without extra explicit training, requiring only \textasciitilde10 A800 days for fine-tuning. Our code is available at: https://github.com/Luo-Yihong/YOSO

340. Trajectory-LLM: A Language-based Data Generator for Trajectory Prediction in Autonomous Driving

链接：https://iclr.cc/virtual/2025/poster/29466 abstract： Vehicle trajectory prediction is a crucial aspect of autonomous driving, which requires extensive trajectory data to train prediction models to understand the complex, varied, and unpredictable patterns of vehicular interactions. However, acquiring real-world data is expensive, so we advocate using Large Language Models (LLMs) to generate abundant and realistic trajectories of interacting vehicles efficiently. These models rely on textual descriptions of vehicle-to-vehicle interactions on a map to produce the trajectories. We introduce Trajectory-LLM (Traj-LLM), a new approach that takes brief descriptions of vehicular interactions as input and generates corresponding trajectories. Unlike language-based approaches that translate text directly to trajectories, Traj-LLM uses reasonable driving behaviors to align the vehicle trajectories with the text. This results in an "interaction-behavior-trajectory" translation process. We have also created a new dataset, Language-to-Trajectory (L2T), which includes 240K textual descriptions of vehicle interactions and behaviors, each paired with corresponding map topologies and vehicle trajectory segments. By leveraging the L2T dataset, Traj-LLM can adapt interactive trajectories to diverse map topologies. Furthermore, Traj-LLM generates additional data that enhances downstream prediction models, leading to consistent performance improvements across public benchmarks. The source code is released at https://github.com/TJU-IDVLab/Traj-LLM.

341. FairMT-Bench: Benchmarking Fairness for Multi-turn Dialogue in Conversational LLMs

链接：https://iclr.cc/virtual/2025/poster/32084 abstract： The increasing deployment of large language model (LLM)-based chatbots has raised concerns regarding fairness. Fairness issues in LLMs may result in serious consequences, such as bias amplification, discrimination, and harm to minority groups. Many efforts are dedicated to evaluating and mitigating biases in LLMs. However, existing fairness benchmarks mainly focus on single-turn dialogues, while multi-turn scenarios, which better reflect real-world conversations, pose greater challenges due to conversational complexity and risk for bias accumulation. In this paper, we introduce a comprehensive benchmark for fairness of LLMs in multi-turn scenarios, FairMT-Bench. Specifically, We propose a task taxonomy to evaluate fairness of LLMs cross three stages: context understanding, interaction fairness, and fairness trade-offs, each comprising two tasks. To ensure coverage of diverse bias types and attributes, our multi-turn dialogue dataset FairMT-10K is constructed by integrating data from established fairness benchmarks. For evaluation, we employ GPT-4 along with bias classifiers like Llama-Guard-3, and human annotators to ensure robustness. Our experiments and analysis on FairMT-10K reveal that in multi-turn dialogue scenarios, LLMs are more prone to generating biased responses, showing significant variation in performance across different tasks and models. Based on these findings, we develop a more challenging dataset, FairMT-1K, and test 15 current state-of-the-art (SOTA) LLMs on this dataset. The results highlight the current state of fairness in LLMs and demonstrate the value of this benchmark for evaluating fairness of LLMs in more realistic multi-turn dialogue contexts. This underscores the need for future works to enhance LLM fairness and incorporate FairMT-1K in such efforts. Our code and dataset are available at https://github.com/FanZT6/FairMT-bench.

342. SynCamMaster: Synchronizing Multi-Camera Video Generation from Diverse Viewpoints

链接：https://iclr.cc/virtual/2025/poster/28481 abstract： Recent advancements in video diffusion models demonstrate remarkable capabilities in simulating real-world dynamics and 3D consistency. This progress motivates us to explore the potential of these models to maintain dynamic consistency across diverse viewpoints, a feature highly sought after in applications like virtual filming. Unlike existing methods focused on multi-view generation of single objects for 4D reconstruction, our interest lies in generating open-world videos from arbitrary viewpoints, incorporating six degrees of freedom (6 DoF) camera poses.To achieve this, we propose a plug-and-play module that enhances a pre-trained text-to-video model for multi-camera video generation, ensuring consistent content across different viewpoints. Specifically, we introduce a multi-view synchronization module designed to maintain appearance and geometry consistency across these viewpoints. Given the scarcity of high-quality training data, we also propose a progressive training scheme that leverages multi-camera images and monocular videos as a supplement to Unreal Engine-rendered multi-camera videos. This comprehensive approach significantly benefits our model.Experimental results demonstrate the superiority of our proposed method over existing competitors and several baselines. Furthermore, our method enables intriguing extensions, such as re-rendering a video from multiple novel viewpoints. Project webpage: https://jianhongbai.github.io/SynCamMaster/

343. PAD: Personalized Alignment of LLMs at Decoding-time

链接：https://iclr.cc/virtual/2025/poster/28948 abstract： Aligning with personalized preferences, which vary significantly across cultural, educational, and political differences, poses a significant challenge due to the computational costs and data demands of traditional alignment methods. In response, this paper presents Personalized Alignment at Decoding-time (PAD), a novel framework designed to align LLM outputs with diverse personalized preferences during the inference phase, eliminating the need for additional training. By introducing a unique personalized reward modeling strategy, this framework decouples the text generation process from personalized preferences, facilitating the generation of generalizable token-level personalized rewards. The PAD algorithm leverages these rewards to guide the decoding process, dynamically tailoring the base model’s predictions to personalized preferences. Extensive experimental results demonstrate that PAD not only outperforms existing training-based

alignment methods in terms of aligning with diverse preferences but also shows significant generalizability to preferences unseen during training and scalability across different base models. This work advances the capability of LLMs to meet user needs in real-time applications, presenting a substantial step forward in personalized LLM alignment.

344. Federated Q-Learning with Reference-Advantage Decomposition: Almost Optimal Regret and Logarithmic Communication Cost

链接：https://iclr.cc/virtual/2025/poster/30326 abstract： In this paper, we consider model-free federated reinforcement learning for tabular episodic Markov decision processes. Under the coordination of a central server, multiple agents collaboratively explore the environment and learn an optimal policy without sharing their raw data. Despite recent advances in federated Q-learning algorithms achieving near-linear regret speedup with low communication cost, existing algorithms only attain suboptimal regrets compared to the information bound. We propose a novel model-free federated Q-Learning algorithm, termed FedQ-Advantage. Our algorithm leverages reference-advantage decomposition for variance reduction and adopts three novel designs: separate event-triggered communication and policy switching, heterogeneous communication triggering conditions, and optional forced synchronization. We prove that our algorithm not only requires a lower logarithmic communication cost but also achieves an almost optimal regret, reaching the information bound up to a logarithmic factor and near-linear regret speedup compared to its single-agent counterpart when the time horizon is sufficiently large.

345. PaCA: Partial Connection Adaptation for Efficient Fine-Tuning

链接：https://iclr.cc/virtual/2025/poster/28689 abstract： Prior parameter-efficient fine-tuning (PEFT) algorithms reduce memory usage and computational costs of fine-tuning large neural network models by training only a few additional adapter parameters, rather than the entire model. However, the reduction in computational costs due to PEFT does not necessarily translate to a reduction in training time; although the computational costs of the adapter layers are much smaller than the pretrained layers, it is well known that those two types of layers are processed sequentially on GPUs, resulting in significant latency overhead. LoRA and its variants avoid this latency overhead by merging the low-rank adapter matrices with the pretrained weights during inference. However, those layers cannot be merged during training since the pretrained weights must remain frozen while the low-rank adapter matrices are updated continuously over the course of training. Furthermore, LoRA and its variants do not reduce activation memory, as the first low-rank adapter matrix still requires the input activations to the pretrained weights to compute weight gradients. To mitigate this issue, we propose Partial Connection Adaptation (PaCA), which fine-tunes randomly selected partial connections within the pretrained weights instead of introducing adapter layers in the model. PaCA not only enhances training speed by eliminating the time overhead due to the sequential processing of the adapter and pretrained layers but also reduces activation memory since only partial activations, rather than full activations, need to be stored for gradient computation. Compared to LoRA, PaCA reduces training time by 22% and total memory usage by 16%, while maintaining comparable accuracy across various fine-tuning scenarios, such as fine-tuning on the MMLU dataset and instruction tuning on the Oasst1 dataset. PaCA can also be combined with quantization, enabling the fine-tuning of large models such as LLaMA3.1-70B. In addition, PaCA enables training with 23% longer sequence and improves throughput by 16\% on both NVIDIA A100 GPU and INTEL Gaudi2 HPU compared to LoRA. The code is available at https://github.com/WooSunghyeon/paca.

346. Gap-Dependent Bounds for Q-Learning using Reference-Advantage Decomposition

链接：https://iclr.cc/virtual/2025/poster/30859 abstract： We study the gap-dependent bounds of two important algorithms for on-policy Q-learning for finite-horizon episodic tabular Markov Decision Processes (MDPs): UCB-Advantage (Zhang et al. 2020) and Q-EarlySettled-Advantage (Li et al. 2021). UCB-Advantage and Q-EarlySettled-Advantage improve upon the results based on Hoeffding-type bonuses and achieve the {almost optimal} $\sqrt{T}$-type regret bound in the worst-case scenario, where T is the total number of steps. However, the benign structures of the MDPs such as a strictly positive suboptimality gap can significantly improve the regret. While gap-dependent regret bounds have been obtained for Q-learning with Hoeffding-type bonuses, it remains an open question to establish gap-dependent regret bounds for Q-learning using variance estimators in their bonuses and reference-advantage decomposition for variance reduction. We develop a novel error decompositionframework to prove gap-dependent regret bounds of UCB-Advantage and Q-EarlySettled-Advantage that are logarithmic in T and improve upon existing ones for Q-learning algorithms. Moreover, we establish the gap-dependent bound for the policy switching cost of UCB-Advantage and improve that under the worst-case MDPs. To our knowledge, this paper presents the first gap-dependent regret analysis for Q-learning using variance estimators and reference-advantage decomposition and also provides the first gap-dependent analysis on policy switching cost for Q-learning.

347. What Matters in Learning from Large-Scale Datasets for Robot Manipulation

链接：https://iclr.cc/virtual/2025/poster/29974 abstract： Imitation learning from large multi-task demonstration datasets has emerged as a promising path for building generally-capable robots. As a result, 1000s of hours have been spent on building such large-scale datasets around the globe. Despite the continuous growth of such efforts, we still lack a systematic understanding of what data should be collected to improve the utility of a robotics dataset and facilitate downstream policy

learning. In this work, we conduct a large-scale dataset composition study to answer this question. We develop a data generation framework to procedurally emulate common sources of diversity in existing datasets (such as sensor placements and object types and arrangements), and use it to generate large-scale robot datasets with controlled compositions, enabling a suite of dataset composition studies that would be prohibitively expensive in the real world. We focus on two practical settings: (1) what types of diversity should be emphasized when future researchers collect large-scale datasets for robotics, and (2) how should current practitioners retrieve relevant demonstrations from existing datasets to maximize downstream policy performance on tasks of interest. Our study yields several critical insights -- for example, we find that camera poses and spatial arrangements are crucial dimensions for both diversity in collection and alignment in retrieval. In real-world robot learning settings, we find that not only do our insights from simulation carry over, but our retrieval strategies on existing datasets such as DROID allow us to consistently outperform existing training strategies by up to 70\%.

348. Accurate and Scalable Graph Neural Networks via Message Invariance

链接：https://iclr.cc/virtual/2025/poster/29451 abstract： Message passing-based graph neural networks (GNNs) have achieved great success in many real-world applications. For a sampled mini-batch of target nodes, the message passing process is divided into two parts: message passing between nodes within the batch (MP-IB) and message passing from nodes outside the batch to those within it (MP-OB). However, MP-OB recursively relies on higher-order out-of-batch neighbors, leading to an exponentially growing computational cost with respect to the number of layers. Due to the neighbor explosion, the whole message passing stores most nodes and edges on the GPU such that many GNNs are infeasible to large-scale graphs. To address this challenge, we propose an accurate and fast mini-batch approach for large graph transductive learning, namely topological compensation (TOP), which obtains the outputs of the whole message passing solely through MP-IB, without the costly MP-OB. The major pillar of TOP is a novel concept of message invariance, which defines message-invariant transformations to convert costly MP-OB into fast MP-IB. This ensures that the modified MP-IB has the same output as the whole message passing. Experiments demonstrate that TOP is significantly faster than existing mini-batch methods by order of magnitude on vast graphs (millions of nodes and billions of edges) with limited accuracy degradation.

349. Wide Neural Networks Trained with Weight Decay Provably Exhibit Neural Collapse

链接：https://iclr.cc/virtual/2025/poster/31212 abstract： Deep neural networks (DNNs) at convergence consistently represent the training data in the last layer via a geometric structure referred to as neural collapse. This empirical evidence has spurred a line of theoretical research aimed at proving the emergence of neural collapse, mostly focusing on the unconstrained features model. Here, the features of the penultimate layer are free variables, which makes the model data-agnostic and puts into question its ability to capture DNN training. Our work addresses the issue, moving away from unconstrained features and studying DNNs that end with at least two linear layers. We first prove generic guarantees on neural collapse that assume \emph{(i)} low training error and balancedness of linear layers (for within-class variability collapse), and \emph{(ii)} bounded conditioning of the features before the linear part (for orthogonality of class-means, and their alignment with weight matrices). The balancedness refers to the fact that $W_{\ell+1}^\top W_{\ell+1}\approx W_\ell W_\ell ^\top$ for any pair ofconsecutive weight matricesof the linear part, and the bounded conditioning requires a well-behaved ratio between largest and smallest non-zero singular values of the features. We then show that such assumptions hold for gradient descent training with weight decay: \emph{(i)} for networks with a wide first layer, we prove low training error and balancedness, and \emph{(ii)} for solutions that are either nearly optimal or stable under large learning rates, we additionally prove the bounded conditioning. Taken together, our results are the first to show neural collapse in the end-to-end training of DNNs.

350. Modeling Fine-Grained Hand-Object Dynamics for Egocentric Video Representation Learning

链接：https://iclr.cc/virtual/2025/poster/29778 abstract： In egocentric video understanding, the motion of hands and objects as well as their interactions play a significant role by nature.However, existing egocentric video representation learning methods mainly focus on aligning video representation with high-level narrations, overlooking the intricate dynamics between hands and objects.In this work, we aim to integrate the modeling of fine-grained hand-object dynamics into the video representation learning process.Since no suitable data is available, we introduce HOD, a novel pipeline employing a hand-object detector and a large language model to generate high-quality narrations with detailed descriptions of hand-object dynamics. To learn these fine-grained dynamics, we propose EgoVideo, a model with a new lightweight motion adapter to capture fine-grained hand-object motion information. Through our co-training strategy, EgoVideo effectively and efficiently leverages the fine-grained hand-object dynamics in the HOD data. Extensive experiments demonstrate that our method achieves state-of-the-art performance across multiple egocentric downstream tasks, including improvements of 6.3% in EK-100 multi-instance retrieval, 5.7% in EK-100 classification, and 16.3% in EGTEA classification in zero-shot settings. Furthermore, our model exhibits robust generalization capabilities in hand-object interaction and robot manipulation tasks.

351. ECHOPulse: ECG Controlled Echocardio-gram Video Generation

链接：https://iclr.cc/virtual/2025/poster/28723 abstract： Echocardiography (ECHO) is essential for cardiac assessments, but its video quality and interpretation heavily relies on manual expertise, leading to inconsistent results from clinical and portable devices. ECHO video generation offers a solution by improving automated monitoring through synthetic data and generating

high-quality videos from routine health data. However, existing models often face high computational costs, slow inference, and rely on complex conditional prompts that require experts' annotations. To address these challenges, we propose ECHOPulse, an ECG-conditioned ECHO video generation model. ECHOPulse introduces two key advancements: (1) it accelerates ECHO video generation by leveraging VQ-VAE tokenization and masked visual token modeling for fast decoding, and (2) it conditions on readily accessible ECG signals, which are highly coherent with ECHO videos, bypassing complex conditional prompts. To the best of our knowledge, this is the first work to use time-series prompts like ECG signals for ECHO video generation. ECHOPulse not only enables controllable synthetic ECHO data generation but also provides updated cardiac function information for disease monitoring and prediction beyond ECG alone. Evaluations on three public and private datasets demonstrate state-of-the-art performance in ECHO video generation across both qualitative and quantitative measures. Additionally, ECHOPulse can be easily generalized to other modality generation tasks, such as cardiac MRI, fMRI, and 3D CT generation. We will make the synthetic ECHO dataset, along with the code and model, publicly available upon acceptance.

352. Complexity Lower Bounds of Adaptive Gradient Algorithms for Non-convex Stochastic Optimization under Relaxed Smoothness

链接：https://iclr.cc/virtual/2025/poster/29203 abstract： Recent results in non-convex stochastic optimization demonstrate the convergence of popular adaptive algorithms (e.g., AdaGrad) under the (L_0, L_1)-smoothness condition, but the rate of convergence is a higher-order polynomial in terms of problem parameters like the smoothness constants. The complexity guaranteed by such algorithms to find an ϵ-stationary point may be significantly larger than the optimal complexity of $\Theta \left(\Delta L \sigma^2 \epsilon^{-4} \right)$ achieved by SGD in the L-smooth setting, where Δ is the initial optimality gap, σ^2 is the variance of stochastic gradient. However, it is currently not known whether these higher-order dependencies can be tightened. To answer this question, we investigate complexity lower bounds for several adaptive optimization algorithms in the (L_0, L_1)-smooth setting, with a focus on the dependence in terms of problem parameters Δ, L_0, L_1. We provide complexity bounds for three variations of AdaGrad, which show at least a quadratic dependence on problem parameters Δ, L_0, L_1. Notably, we show that the decorrelated variant of AdaGrad-Norm requires at least $\Omega \left(\Delta^2 L_1^2 \sigma^2 \epsilon^{-4} \right)$ stochastic gradient queries to find an ϵ-stationary point. We also provide a lower bound for SGD with a broad class of adaptive stepsizes. Our results show that, for certain adaptive algorithms, the (L_0, L_1)-smooth setting is fundamentally more difficult than the standard smooth setting, in terms of the initial optimality gap and the smoothness constants.

353. Local Steps Speed Up Local GD for Heterogeneous Distributed Logistic Regression

链接：https://iclr.cc/virtual/2025/poster/28491 abstract： We analyze two variants of Local Gradient Descent applied to distributed logistic regression with heterogeneous, separable data and show convergence at the rate $O(1/KR)$ for K local steps and sufficiently large R communication rounds. In contrast, all existing convergence guarantees for Local GD applied to any problem are at least $\Omega(1/R)$, meaning they fail to show the benefit of local updates. The key to our improved guarantee is showing progress on the logistic regression objective when using a large stepsize $\eta \gg 1/K$, whereas prior analysis depends on $\eta \leq 1/K$.

354. PuzzleFusion++: Auto-agglomerative 3D Fracture Assembly by Denoise and Verify

链接：https://iclr.cc/virtual/2025/poster/30838 abstract： This paper proposes a novel “auto-agglomerative” 3D fracture assembly method, PuzzleFusion++, resembling how humans solve challenging spatial puzzles. Starting from individual fragments, the approach 1) aligns and merges fragments into larger groups akin to agglomerative clustering and 2) repeats the process iteratively in completing the assembly akin to auto-regressive methods. Concretely, a diffusion model denoises the 6-DoF alignment parameters of the fragments simultaneously,and a transformer model verifies and merges pairwise alignments into larger ones, whose process repeats iteratively. Extensive experiments on the Breaking Bad dataset show that PuzzleFusion++ outperforms all other state-of-the-art techniques by significant margins across all metrics In particular by over 10% in part accuracy and 50% in Chamfer distance. We will release code and model.

355. MEGA-Bench: Scaling Multimodal Evaluation to over 500 Real-World Tasks

链接：https://iclr.cc/virtual/2025/poster/31110 abstract： We present MEGA-Bench, an evaluation suite that scales multimodal evaluation to over 500 real-world tasks, to address the highly heterogeneous daily use cases of end users.Our objective is to optimize for a set of high-quality data samples that cover a highly diverse and rich set of multimodal tasks, while enabling cost-effective and accurate model evaluation.In particular, we collected 505 realistic tasks encompassing over 8,000 samples from 16 expert annotators to extensively cover the multimodal task space. Instead of unifying these problems into standard multi-choice questions (like MMMU, MM-Bench, and MMT-Bench), we embrace a wide range of output formats like numbers, phrases, code, \LaTeX, coordinates, JSON, free-form, etc. To accommodate these formats, we developed over 40 metrics to evaluate these tasks. Unlike existing benchmarks, MEGA-Bench offers a fine-grained capability report across multiple dimensions (e.g., application, input type, output format, skill), allowing users to interact with and visualize model capabilities in

depth. We evaluate a wide variety of frontier vision-language models on MEGA-Bench to understand their capabilities across these dimensions.

356. SparsyFed: Sparse Adaptive Federated Learning

链接：https://iclr.cc/virtual/2025/poster/29840 abstract： Sparse training is often adopted in cross-device federated learning (FL) environments where constrained devices collaboratively train a machine learning model on private data by exchanging pseudo-gradients across heterogeneous networks. Although sparse training methods can reduce communication overhead and computational burden in FL, they are often not used in practice for the following key reasons: (1) data heterogeneity makes it harder for clients to reach consensus on sparse models compared to dense ones, requiring longer training; (2) methods for obtaining sparse masks lack adaptivity to accommodate very heterogeneous data distributions, crucial in cross-device FL; and (3) additional hyperparameters are required, which are notably challenging to tune in FL. This paper presents SparsyFed, a practical federated sparse training method that critically addresses the problems above. Previous works have only solved one or two of these challenges at the expense of introducing new trade-offs, such as clients' consensus on masks versus sparsity pattern adaptivity. We show that SparsyFed simultaneously (1) can produce 95% sparse models, with negligible degradation in accuracy, while only needing a single hyperparameter, (2) achieves a per-round weight regrowth 200 times smaller than previous methods, and (3) allows the sparse masks to adapt to highly heterogeneous data distributions and outperform all baselines under such conditions.

357. CG-Bench: Clue-grounded Question Answering Benchmark for Long Video Understanding

链接：https://iclr.cc/virtual/2025/poster/28509 abstract： The existing video understanding benchmarks for multimodal large language models (MLLMs) mainly focus on short videos. The few benchmarks for long video understanding often rely on multiple-choice questions (MCQs). Due to the limitations of MCQ evaluations and the advanced reasoning abilities of MLLMs, models can often answer correctly by combining short video insights with elimination, without truly understanding the content. To bridge this gap, we introduce CG-Bench, a benchmark for clue-grounded question answering in long videos. CG-Bench emphasizes the model's ability to retrieve relevant clues, enhancing evaluation credibility. It includes 1,219 manually curated videos organized into 14 primary, 171 secondary, and 638 tertiary categories, making it the largest benchmark for long video analysis. The dataset features 12,129 QA pairs in three question types: perception, reasoning, and hallucination. To address the limitations of MCQ-based evaluation, we develop two novel clue-based methods: clue-grounded white box and black box evaluations, assessing whether models generate answers based on accurate video understanding. We evaluated multiple closed-source and open-source MLLMs on CG-Bench. The results show that current models struggle significantly with long videos compared to short ones, and there is a notable gap between open-source and commercial models. We hope CG-Bench will drive the development of more reliable and capable MLLMs for long video comprehension.

358. Bootstrapping Language-Guided Navigation Learning with Self-Refining Data Flywheel

链接：https://iclr.cc/virtual/2025/poster/29824 abstract： Creating high-quality data for training robust language-instructed agents is a long-lasting challenge in embodied AI. In this paper, we introduce a Self-Refining Data Flywheel (SRDF) that generates high-quality and large-scale navigational instruction-trajectory pairs by iteratively refining the data pool through the collaboration between two models, the instruction generator and the navigator, without any human-in-the-loop annotation. Specifically, SRDF starts with using a base generator to create an initial data pool for training a base navigator, followed by applying the trained navigator to filter the data pool. This leads to higher-fidelity data to train a better generator, which can, in turn, produce higher-quality data for training the next-round navigator. Such a flywheel establishes a data self-refining process, yielding a continuously improved and highly effective dataset for large-scale language-guided navigation learning. Our experiments demonstrate that after several flywheel rounds, the navigator elevates the performance boundary from 70\% to 78\% SPL on the classic R2R test set, surpassing human performance (76\%) for the first time. Meanwhile, this process results in a superior generator, evidenced by a SPICE increase from 23.5 to 26.2, better than all previous VLN instruction generation methods. Finally, we demonstrate the scalability of our method through increasing environment and instruction diversity, andthe generalization ability of our pre-trained navigator across various downstream navigation tasks, surpassing state-of-the-art methods by a large margin in all cases.

359. Instruct-SkillMix: A Powerful Pipeline for LLM Instruction Tuning

链接：https://iclr.cc/virtual/2025/poster/31037 abstract： We introduce INSTRUCT-SKILLMIX, an automated approach for creating diverse, high quality SFT data for instruction-following. The pipeline involves two stages, each leveraging an existing powerful LLM: (1) Skill extraction: uses the LLM to extract core "skills" for instruction-following by directly prompting the model. This is inspired by "LLM metacognition" of (Didolkar et al., 2024); (2) Data generation: uses the powerful LLM to generate (instruction, response) data thatexhibit a randomly chosen pair of these skills. Here, the use of random skill combinations promotes diversity and difficulty. The estimated cost of creating the dataset is under $600. Vanilla SFT (i.e., no PPO, DPO, or RL methods) on data generated from INSTRUCT-SKILLMIX leads to strong gains on instruction following benchmarks such as AlpacaEval 2.0, MT-Bench, and WildBench. With just 4K examples, LLaMA-3-8B-Base achieves 42.76% length-controlled win rate on AlpacaEval 2.0, a level similar to frontier models like Claude 3 Opus and LLaMA-3.1-405B-Instruct. Ablation studies also

suggest plausible reasons for why creating open instruction-tuning datasets via naive crowd-sourcing has proved difficult. In our dataset,adding 20% low quality answers (“shirkers”) causes a noticeable degradation in performance.The INSTRUCT-SKILLMIX pipeline seems flexible and adaptable to other settings.

360. Robust-PIFu: Robust Pixel-aligned Implicit Function for 3D Human Digitalization from a Single Image

链接：https://iclr.cc/virtual/2025/poster/28845 abstract： Existing methods for 3D clothed human digitalization perform well when the input image is captured in ideal conditions that assume the lack of any occlusion. However, in reality, images may often have occlusion problems such as incomplete observation of the human subject's full body, self-occlusion by the human subject, and non-frontal body pose. When given such input images, these existing methods fail to perform adequately. Thus, we propose Robust-PIFu, a pixel-aligned implicit model that capitalized on large-scale, pretrained latent diffusion models to address the challenge of digitalizing human subjects from non-ideal images that suffer from occlusions.Robust-PIfu offers four new contributions. Firstly, we propose a 'disentangling' latent diffusion model. This diffusion model, pretrained on billions of images, takes in any input image and removes external occlusions, such as inter-person occlusions, from that image. Secondly, Robust-PIFu addresses internal occlusions like self-occlusion by introducing a `penetrating' latent diffusion model. This diffusion model outputs multi-layered normal maps that by-pass occlusions caused by the human subject's own limbs or other body parts (i.e. self-occlusion). Thirdly, in order to incorporate such multi-layered normal maps into a pixel-aligned implicit model, we introduce our Layered-Normals Pixel-aligned Implicit Model, which improves the structural accuracy of predicted clothed human meshes. Lastly, Robust-PIFu proposes an optional super-resolution mechanism for the multi-layered normal maps. This addresses scenarios where the input image is of low or inadequate resolution. Though not strictly related to occlusion, this is still an important subproblem. Our experiments show that Robust-PIFu outperforms current SOTA methods both qualitatively and quantitatively. Our code will be released to the public.

361. Exact Computation of Any-Order Shapley Interactions for Graph Neural Networks

链接：https://iclr.cc/virtual/2025/poster/30660 abstract： Albeit the ubiquitous use of Graph Neural Networks (GNNs) in machine learning (ML) prediction tasks involving graph-structured data, their interpretability remains challenging. In explainable artificial intelligence (XAI), the Shapley Value (SV) is the predominant method to quantify contributions of individual features to a ML model’s output. Addressing the limitations of SVs in complex prediction models, Shapley Interactions (SIs) extend the SV to groups of features. In this work, we explain single graph predictions of GNNs with SIs that quantify node contributions and interactions among multiple nodes. By exploiting the GNN architecture, we show that the structure of interactions in node embeddings are preserved for graph prediction. As a result, the exponential complexity of SIs depends only on the receptive fields, i.e. the message-passing ranges determined by the connectivity of the graph and the number of convolutional layers. Based on our theoretical results, we introduce GraphSHAP-IQ, an efficient approach to compute any-order SIs exactly. GraphSHAP-IQ is applicable to popular message passing techniques in conjunction with a linear global pooling and output layer. We showcase that GraphSHAP-IQ substantially reduces the exponential complexity of computing exact SIs on multiple benchmark datasets. Beyond exact computation, we evaluate GraphSHAP-IQ’s approximation of SIs on popular GNN architectures and compare with existing baselines. Lastly, we visualize SIs of real-world water distribution networks and molecule structures using a SI-Graph.

362. Probe Pruning: Accelerating LLMs through Dynamic Pruning via Model-Probing

链接：https://iclr.cc/virtual/2025/poster/29365 abstract： We introduce Probe Pruning (PP), a novel framework for online, dynamic, structured pruning of Large Language Models (LLMs) applied in a batch-wise manner. PP leverages the insight that not all samples and tokens contribute equally to the model's output, and probing a small portion of each batch effectively identifies crucial weights, enabling tailored dynamic pruning for different batches. It comprises three main stages: probing, history-informed pruning, and full inference. In the probing stage, PP selects a small yet crucial set of hidden states, based on residual importance, to run a few model layers ahead. During the history-informed pruning stage, PP strategically integrates the probing states with historical states. Subsequently, it structurally prunes weights based on the integrated states and the PP importance score, a metric developed specifically to assess the importance of each weight channel in maintaining performance. In the final stage, full inference is conducted on the remaining weights. A major advantage of PP is its compatibility with existing models, as it operates without requiring additional neural network modules or fine-tuning. Comprehensive evaluations of PP on LLaMA-2/3 and OPT models reveal that even minimal probing—using just 1.5% of FLOPs—can substantially enhance the efficiency of structured pruning of LLMs. For instance, when evaluated on LLaMA-2-7B with WikiText2, PP achieves a 2.56 times lower ratio of performance degradation per unit of latency reduction compared to the state-of-the-art method at a 40\% pruning ratio.

363. Discrete Distribution Networks

链接：https://iclr.cc/virtual/2025/poster/27788 abstract： We introduce a novel generative model, the Discrete Distribution Networks (DDN), that approximates data distribution using hierarchical discrete distributions. We posit that since the features

within a network inherently capture distributional information, enabling the network to generate multiple samples simultaneously, rather than a single output, may offer an effective way to represent distributions. Therefore, DDN fits the target distribution, including continuous ones, by generating multiple discrete sample points. To capture finer details of the target data, DDN selects the output that is closest to the Ground Truth (GT) from the coarse results generated in the first layer. This selected output is then fed back into the network as a condition for the second layer, thereby generating new outputs more similar to the GT. As the number of DDN layers increases, the representational space of the outputs expands exponentially, and the generated samples become increasingly similar to the GT. This hierarchical output pattern of discrete distributions endows DDN with unique properties: more general zero-shot conditional generation and 1D latent representation. We demonstrate the efficacy of DDN and its intriguing properties through experiments on CIFAR-10 and FFHQ. The code is available at https://discrete-distribution-networks.github.io/

364. The Unreasonable Ineffectiveness of the Deeper Layers

链接：https://iclr.cc/virtual/2025/poster/28400 abstract： How is knowledge stored in an LLM's weights? We study this via layer pruning: if removing a certain layer does not affect model performance in common question-answering benchmarks, then the weights in that layer are not necessary for storing the knowledge needed to answer those questions. To find these unnecessary parameters, we identify the optimal block of layers to prune by considering similarity across layers; then, to "heal" the damage, we perform a small amount of finetuning. Surprisingly, with this method we find minimal degradation of performance until after a large fraction (up to half) of the layers are removed for some common open-weight models. From a scientific perspective, the robustness of these LLMs to the deletion of layers implies either that current pretraining methods are not properly leveraging the parameters in the deeper layers of the network or that the shallow layers play a critical role in storing knowledge. For our study, we use parameter-efficient finetuning (PEFT) methods, specifically quantization and Low Rank Adapters (QLoRA), such that each of our experiments can be performed on a single 40GB A100 GPU.

365. Divergence-enhanced Knowledge-guided Context Optimization for Visual-Language Prompt Tuning

链接：https://iclr.cc/virtual/2025/poster/30858 abstract： Prompt tuning vision-language models like CLIP has shown great potential in learning transferable representations for various downstream tasks. The main issue is how to mitigate the over-fitting problem on downstream tasks with limited training samples. While knowledge-guided context optimization has been proposed by constructing consistency constraints to handle catastrophic forgetting in the pre-trained backbone, it also introduces a bias toward pre-training. This paper proposes a novel and simple Divergence-enhanced Knowledge-guided Prompt Tuning (DeKg) method to address this issue.The key insight is that the bias toward pre-training can be alleviated by encouraging the independence between the learnable and the crafted prompt. Specifically, DeKg employs the Hilbert-Schmidt Independence Criterion (HSIC) to regularize the learnable prompts, thereby reducing their dependence on prior general knowledge, and enabling divergence induced by target knowledge. Comprehensive evaluations demonstrate that DeKg serves as a plug-and-play module can seamlessly integrate with existing knowledge-guided context optimization methods and achieves superior performance in three challenging benchmarks. We make our code available at https://github.com/cnunlp/DeKg.

366. Physics-Informed Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/28020 abstract： Generative models such as denoising diffusion models are quickly advancing their ability to approximate highly complex data distributions. They are also increasingly leveraged in scientific machine learning, where samples from the implied data distribution are expected to adhere to specific governing equations. We present a framework that unifies generative modeling and partial differential equation fulfillment by introducing a first-principle-based loss term that enforces generated samples to fulfill the underlying physical constraints. Our approach reduces the residual error by up to two orders of magnitudes compared to previous work in a fluid flow case study and outperforms task-specific frameworks in relevant metrics for structural topology optimization. We also present numerical evidence that our extended training objective acts as a natural regularization mechanism against overfitting. Our framework is simple to implement and versatile in its applicability for imposing equality and inequality constraints as well as auxiliary optimization objectives. Code is available at https://github.com/jhbastek/PhysicsInformedDiffusionModels.

367. Adaptive Methods through the Lens of SDEs: Theoretical Insights on the Role of Noise

链接：https://iclr.cc/virtual/2025/poster/27812 abstract： Despite the vast empirical evidence supporting the efficacy of adaptive optimization methods in deep learning, their theoretical understanding is far from complete. This work introduces novel SDEs for commonly used adaptive optimizers: SignSGD, RMSprop(W), and Adam(W). These SDEs offer a quantitatively accurate description of these optimizers and help illuminate an intricate relationship between adaptivity, gradient noise, and curvature. Our novel analysis of SignSGD highlights a noteworthy and precise contrast to SGD in terms of convergence speed, stationary distribution, and robustness to heavy-tail noise. We extend this analysis to AdamW and RMSpropW, for which we observe that the role of noise is much more complex. Crucially, we support our theoretical analysis with experimental evidence by verifying our insights: this includes numerically integrating our SDEs using Euler-Maruyama discretization on various neural network architectures such as MLPs, CNNs, ResNets, and Transformers. Our SDEs accurately track the behavior of the respective optimizers, especially when compared to previous SDEs derived for Adam and RMSprop. We believe our approach

can provide valuable insights into best training practices and novel scaling rules.

368. Functional Homotopy: Smoothing Discrete Optimization via Continuous Parameters for LLM Jailbreak Attacks

链接：https://iclr.cc/virtual/2025/poster/27963 abstract： Optimization methods are widely employed in deep learning to address and mitigate undesired model responses. While gradient-based techniques have proven effective for image models, their application to language models is hindered by the discrete nature of the input space. This study introduces a novel optimization approach, termed the functional homotopy method, which leverages the functional duality between model training and input generation. By constructing a series of easy-to-hard optimization problems, we iteratively solve these using principles derived from established homotopy methods. We apply this approach to jailbreak attack synthesis for large language models (LLMs), achieving a 20%-30% improvement in success rate over existing methods in circumventing established safe open-source models such as Llama-2 and Llama-3.

369. Predictive Uncertainty Quantification for Bird's Eye View Segmentation: A Benchmark and Novel Loss Function

链接：https://iclr.cc/virtual/2025/poster/28599 abstract： The fusion of raw sensor data to create a Bird's Eye View (BEV) representation is critical for autonomous vehicle planning and control. Despite the growing interest in using deep learning models for BEV semantic segmentation, anticipating segmentation errors and enhancing the explainability of these models remain underexplored. This paper introduces a comprehensive benchmark for predictive uncertainty quantification in BEV segmentation, evaluating multiple uncertainty quantification methods across three popular datasets with three representative network architectures. Our study focuses on the effectiveness of quantified uncertainty in detecting misclassified and out-of-distribution (OOD) pixels while also improving model calibration. Through empirical analysis, we uncover challenges in existing uncertainty quantification methods and demonstrate the potential of evidential deep learning techniques, which capture both aleatoric and epistemic uncertainty. To address these challenges, we propose a novel loss function, Uncertainty-Focal-Cross-Entropy (UFCE), specifically designed for highly imbalanced data, along with a simple uncertainty-scaling regularization term that improves both uncertainty quantification and model calibration for BEV segmentation.

370. Adapt-∞: Scalable Continual Multimodal Instruction Tuning via Dynamic Data Selection

链接：https://iclr.cc/virtual/2025/poster/30379 abstract： Visual instruction datasets from various distributors are released at different times and often contain a significant number of semantically redundant text-image pairs, depending on their task compositions (i.e., skills) or reference sources. This redundancy greatly limits the efficient deployment of continually adaptable multimodal large language models, hindering their ability to refine existing skills and acquire new competencies over time. To address this, we reframe the problem of lifelong Instruction Tuning (LiIT) via data selection, where the model automatically selects beneficial samples to learn from earlier and new datasets based on the current state of acquired knowledge in the model. Based on empirical analyses that show that selecting the best data subset using a static importance measure is often ineffective for multi-task datasets with evolving distributions, we propose Adapt-∞, a new multi-way and adaptive data selection approach that dynamically balances sample efficiency and effectiveness during LiIT. We first construct pseudo-skill clusters by grouping gradient-based sample vectors. Next, we select the best-performing data selector for each skill cluster from a pool of selector experts, including our newly proposed scoring function, Image Grounding score. This data selector samples a subset of the most important samples from each skill cluster for training. To prevent the continuous increase in the size of the dataset pool during LIT, which would result in excessive computation, we further introduce a cluster-wise permanent data pruning strategy to remove the most semantically redundant samples from each cluster, keeping computational requirements manageable. We validate the effectiveness and efficiency of Adapt-∞ over a sequence of various multimodal instruction tuning datasets with various tasks, including (Knowledge) VQA, multilingual, grounding, reasoning, language-only, and multi-image comprehension tasks. Training with samples selected by Adapt-∞ alleviates catastrophic forgetting, especially for rare tasks, and promotes forward transfer across the continuum using only a fraction of the original datasets.

371. econSG: Efficient and Multi-view Consistent Open-Vocabulary 3D Semantic Gaussians

链接：https://iclr.cc/virtual/2025/poster/28248 abstract： The primary focus of most recent works on open-vocabulary neural fields is extracting precise semantic featuresfrom the VLMs and then consolidating them efficiently into a multi-view consistent 3D neural fieldsrepresentation. However, most existing works over-trusted SAM to regularize image-level CLIP without any further refinement. Moreover, several existing works improved efficiency by dimensionality reduction of semantic features from 2D VLMs before fusing with 3DGS semantic fields, which inevitably leads to multi-view inconsistency. In this work, we propose econSG for open-vocabulary semantic segmentation with 3DGS. Our econSG consists of: 1) A Confidence-region Guided Regularization (CRR) that mutually refines SAM and CLIP to get the best of both worlds for precise semantic features with complete and precise boundaries. 2) A low dimensional contextual space to enforce 3D multi-view consistency while improving computational efficiency by fusing backprojected multi-view 2D features and follow by dimensional reduction directly on the fused 3D features instead of operating on each 2D view separately. Our econSG show state-of-the-art performance on four

benchmark datasets compared to the existing methods. Furthermore, we are also the most efficient training among all the methods.

372. OmniCorpus: A Unified Multimodal Corpus of 10 Billion-Level Images Interleaved with Text

链接：https://iclr.cc/virtual/2025/poster/32063 abstract： Image-text interleaved data, consisting of multiple images and texts arranged in a natural document format, aligns with the presentation paradigm of internet data and closely resembles human reading habits. Recent studies have shown that such data aids multimodal in-context learning and maintains the capabilities of large language models during multimodal fine-tuning. However, the limited scale and diversity of current image-text interleaved data restrict the development of multimodal large language models. In this paper, we introduce OmniCorpus, a 10 billion-scale image-text interleaved dataset. Using an efficient data engine, we filter and extract large-scale high-quality documents, which contain 8.6 billion images and 1,696 billion text tokens. Compared to counterparts (e.g., MMC4, OBELICS), our dataset 1) has 15 times larger scales while maintaining good data quality; 2) features more diverse sources, including both English and non-English websites as well as video-centric websites; 3) is more flexible, easily degradable from an image-text interleaved format to pure text corpus and image-text pairs. Through comprehensive analysis and experiments, we validate the quality, usability, and effectiveness of the proposed dataset. We hope this could provide a solid data foundation for future multimodal model research.

373. RaSA: Rank-Sharing Low-Rank Adaptation

链接：https://iclr.cc/virtual/2025/poster/30271 abstract： Low-rank adaptation (LoRA) has been prominently employed for parameter-efficient fine-tuning of large language models (LLMs). However, the limited expressive capacity of LoRA, stemming from the low-rank constraint, has been recognized as a bottleneck, particularly in rigorous tasks like code generation and mathematical reasoning. To address this limitation, we introduce Rank-Sharing Low-Rank Adaptation (RaSA), an innovative extension that enhances the expressive capacity of LoRA by leveraging partial rank sharing across layers. By forming a shared rank pool and applying layer-specific weighting, RaSA effectively increases the number of ranks without augmenting parameter overhead. Our theoretically grounded and empirically validated approach demonstrates that RaSA not only maintains the core advantages of LoRA but also significantly boosts performance in challenging code and math tasks. Code, data and scripts are available at: https://github.com/zwhe99/RaSA.

374. MLPs Learn In-Context on Regression and Classification Tasks

链接：https://iclr.cc/virtual/2025/poster/29924 abstract： In-context learning (ICL), the remarkable ability to solve a task from only input exemplars, is often assumed to be a unique hallmark of Transformer models. By examining commonly employed synthetic ICL tasks, we demonstrate that multi-layer perceptrons (MLPs) can also learn in-context. Moreover, MLPs, and the closely related MLP-Mixer models, learn in-context comparably with Transformers under the same compute budget in this setting. We further show that MLPs outperform Transformers on a series of classical tasks from psychology designed to test relational reasoning, which are closely related to in-context classification. These results underscore a need for studying in-context learning beyond attention-based architectures, while also challenging prior arguments against MLPs' ability to solve relational tasks. Altogether, our results highlight the unexpected competence of MLPs in a synthetic setting, and support the growing interest in all-MLP alternatives to Transformer architectures. It remains unclear how MLPs perform against Transformers at scale on real-world tasks, and where a performance gap may originate. We encourage further exploration of these architectures in more complex settings to better understand the potential comparative advantage of attention-based schemes.

375. Re-Evaluating the Impact of Unseen-Class Unlabeled Data on Semi-Supervised Learning Model

链接：https://iclr.cc/virtual/2025/poster/29363 abstract： Semi-supervised learning (SSL) effectively leverages unlabeled data and has been proven successful across various fields. Current safe SSL methods believe that unseen classes in unlabeled data harm the performance of SSL models. However, previous methods for assessing the impact of unseen classes on SSL model performance are flawed. They fix the size of the unlabeled dataset and adjust the proportion of unseen classes within the unlabeled data to assess the impact. This process contravenes the principle of controlling variables. Adjusting the proportion of unseen classes in unlabeled data alters the proportion of seen classes, meaning the decreased classification performance of seen classes may not be due to an increase in unseen class samples in the unlabeled data, but rather a decrease in seen class samples. Thus, the prior flawed assessment standard that "unseen classes in unlabeled data can damage SSL model performance" may not always hold true. This paper strictly adheres to the principle of controlling variables, maintaining the proportion of seen classes in unlabeled data while only changing the unseen classes across five critical dimensions, to investigate their impact on SSL models from global robustness and local robustness. Experiments demonstrate that unseen classes in unlabeled data do not necessarily impair the performance of SSL models; in fact, under certain conditions, unseen classes may even enhance them.

376. GOAL: A Generalist Combinatorial Optimization Agent Learner

链接：https://iclr.cc/virtual/2025/poster/27692 abstract： Machine Learning-based heuristics have recently shown impressive performance in solving a variety of hard combinatorial optimization problems (COPs). However they generally rely on a separate neural model, specialized and trained for each single problem. Any variation of a problem requires adjustment of its model and re-training from scratch. In this paper, we propose GOAL (for Generalist combinatorial Optimization Agent Learner), a generalist model capable of efficiently solving multiple COPs and which can be fine-tuned to solve new COPs. GOAL consists of a single backbone plus light-weight problem-specific adapters for input and output processing. The backbone is based on a new form of mixed-attention blocks which allows to handle problems defined on graphs with arbitrary combinations of node, edge and instance-level features. Additionally, problems which involve heterogeneous types of nodes or edges are handled through a novel multi-type transformer architecture, where the attention blocks are duplicated to attend the meaningful combinations of types while relying on the same shared parameters. We train GOAL on a set of routing, scheduling and classic graph problems and show that it is only slightly inferior to the specialized baselines while being the first multi-task model that solves a wide range of COPs. Finally we showcase the strong transfer learning capacity of GOAL by fine-tuning it on several new problems. Our code is available at https://github.com/naver/goal-co .

377. Do Vision & Language Decoders use Images and Text equally? How Self-consistent are their Explanations?

链接：https://iclr.cc/virtual/2025/poster/28530 abstract： Vision and language model (VLM) decoders are currently the best-performing architectures on multimodal tasks. Next to answers, they are able to produce natural language explanations, either in post-hoc or CoT settings. However, it is not clear to what extent they are using the input vision and text modalities when generating answers or explanations.In this work, we investigate if VLMs rely on their input modalities differently when they produce explanations as opposed to answers.We also evaluate the self-consistency of VLM decoders in both post-hoc and CoT explanation settings, by extending existing unimodal tests and measures to VLM decoders. We find that most tested VLMs are less self-consistent than LLMs. Text contributions in all tested VL decoders are more important than image contributions in all examined tasks. However, when comparing explanation generation to answer generation, the contributions of images are significantly stronger for generating explanations compared to answers. This difference is even larger in CoT compared to post-hoc explanations.Lastly, we provide an up-to-date benchmarking of state-of-the-art VL decoders on the VALSE benchmark, which before was restricted to VL encoders. We find that the tested VL decoders still struggle with most phenomena tested by VALSE. We will make our code publicly available.

378. Towards Federated RLHF with Aggregated Client Preference for LLMs

链接：https://iclr.cc/virtual/2025/poster/28445 abstract： Reinforcement learning with human feedback (RLHF) fine-tunes a pretrained large language model (LLM) using user preference data, enabling it to generate content aligned with human preferences. However, due to privacy concerns, users may be reluctant to share sensitive preference data. To address this, we propose utilizing Federated Learning (FL) techniques, allowing large-scale preference collection from diverse real-world users without requiring them to transmit data to a central server. Our federated RLHF methods (i.e., FedBis and FedBiscuit) encode each client's preferences into binary selectors and aggregate them to capture common preferences. In particular, FedBiscuit overcomes key challenges, such as preference heterogeneity and reward hacking, through innovative solutions like grouping clients with similar preferences to reduce heterogeneity and using multiple binary selectors to enhance LLM output quality. To evaluate the performance of the proposed methods, we establish the first federated RLHF benchmark with a heterogeneous human preference dataset. Experimental results show that by integrating the LLM with aggregated client preferences, FedBis and FedBiscuit significantly enhance the professionalism and readability of the generated content.

379. Weak-to-Strong Preference Optimization: Stealing Reward from Weak Aligned Model

链接：https://iclr.cc/virtual/2025/poster/28894 abstract： Aligning language models (LMs) with human preferences has become a key area of research, enabling these models to meet diverse user needs better. Inspired by weak-to-strong generalization, where a strong LM fine-tuned on labels generated by a weaker model can consistently outperform its weak supervisor, we extend this idea to model alignment. In this work, we observe that the alignment behavior in weaker models can be effectively transferred to stronger models and even exhibit an amplification effect. Based on this insight, we propose a method called Weak-to-Strong Preference Optimization (WSPO), which achieves strong model alignment by learning the distribution differences before and after the alignment of the weak model. Experiments demonstrate that WSPO delivers outstanding performance, improving the win rate of Qwen2-7B-Instruct on Arena-Hard from 39.70 to 49.60, achieving a remarkable 47.04 length-controlled win rate on AlpacaEval 2, and scoring 7.33 on MT-bench. Our results suggest that using the weak model to elicit a strong model with a high alignment ability is feasible. The code is available at https://github.com/zwhong714/weak-to-strong-preference-optimization.

380. Do Large Language Models Truly Understand Geometric Structures?

链接：https://iclr.cc/virtual/2025/poster/30330 abstract： Geometric ability is a significant challenge for large language models (LLMs) due to the need for advanced spatial comprehension and abstract thinking. Existing datasets primarily evaluate LLMs on their final answers, but they cannot truly measure their true understanding of geometric structures, as LLMs can arrive at correct answers by coincidence. To fill this gap, we introduce the GeomRel dataset, designed to evaluate LLMs' understanding of

geometric structures by isolating the core step of geometric relationship identification in problem-solving. Using this benchmark, we conduct thorough evaluations of diverse LLMs and identify key limitations in understanding geometric structures. We further propose the Geometry Chain-of-Thought (GeoCoT) method, which enhances LLMs' ability to identify geometric relationships, resulting in significant performance improvements.

381. FreqPrior: Improving Video Diffusion Models with Frequency Filtering Gaussian Noise

链接：https://iclr.cc/virtual/2025/poster/30733 abstract： Text-driven video generation has advanced significantly due to developments in diffusion models. Beyond the training and sampling phases, recent studies have investigated noise priors of diffusion models, as improved noise priors yield better generation results. One recent approach employs the Fourier transform to manipulate noise, marking the initial exploration of frequency operations in this context. However, it often generates videos that lack motion dynamics and imaging details. In this work, we provide a comprehensive theoretical analysis of the variance decay issue present in existing methods, contributing to the loss of details and motion dynamics. Recognizing the critical impact of noise distribution on generation quality, we introduce FreqPrior, a novel noise initialization strategy that refines noise in the frequency domain. Our method features a novel filtering technique designed to address different frequency signals while maintaining the noise prior distribution that closely approximates a standard Gaussian distribution. Additionally, we propose a partial sampling process by perturbing the latent at an intermediate timestep while finding the noise prior, significantly reducing inference time without compromising quality. Extensive experiments on VBench demonstrate that our method achieves the highest scores in both quality and semantic assessments, resulting in the best overall total score. These results highlight the superiority of our proposed noise prior.

382. Endowing Visual Reprogramming with Adversarial Robustness

链接：https://iclr.cc/virtual/2025/poster/29796 abstract： Visual reprogramming (VR) leverages well-developed pre-trained models (e.g., a pre-trained classifier on ImageNet) to tackle target tasks (e.g., a traffic sign recognition task), without the need for training from scratch. Despite the effectiveness of previous VR methods, all of them did not consider the adversarial robustness of reprogrammed models against adversarial attacks, which could lead to unpredictable problems in safety-crucial target tasks. In this paper, we empirically find that reprogramming pre-trained models with adversarial robustness and incorporating adversarial samples from the target task during reprogramming can both improve the adversarial robustness of reprogrammed models. Furthermore, we propose a theoretically guaranteed adversarial robustness risk upper bound for VR, which validates our empirical findings and could provide a theoretical foundation for future research. Extensive experiments demonstrate that by adopting the strategies revealed in our empirical findings, the adversarial robustness of reprogrammed models can be enhanced.

383. ORSO: Accelerating Reward Design via Online Reward Selection and Policy Optimization

链接：https://iclr.cc/virtual/2025/poster/31232 abstract： Reward shaping is critical in reinforcement learning (RL), particularly for complex tasks where sparse rewards can hinder learning. However, choosing effective shaping rewards from a set of reward functions in a computationally efficient manner remains an open challenge. We propose Online Reward Selection and Policy Optimization (ORSO), a novel approach that frames the selection of shaping reward function as an online model selection problem. ORSO automatically identifies performant shaping reward functions without human intervention with provable regret guarantees. We demonstrate ORSO's effectiveness across various continuous control tasks. Compared to prior approaches, ORSO significantly reduces the amount of data required to evaluate a shaping reward function, resulting in superior data efficiency and a significant reduction in computational time (up to 8×). ORSO consistently identifies high-quality reward functions outperforming prior methods by more than 50% and on average identifies policies as performant as the ones learned using manually engineered reward functions by domain experts.

384. On the Crucial Role of Initialization for Matrix Factorization

链接：https://iclr.cc/virtual/2025/poster/29262 abstract： This work revisits the classical low-rank matrix factorization problem and unveils the critical role of initialization in shaping convergence rates for such nonconvex and nonsmooth optimization. We introduce Nystrom initialization, which significantly improves the global convergence of Scaled Gradient Descent (ScaledGD) in both symmetric and asymmetric matrix factorization tasks. Specifically, we prove that ScaledGD with Nystrom initialization achieves quadratic convergence in cases where only linear rates were previously known. Furthermore, we extend this initialization to low-rank adapters (LoRA) commonly used for finetuning foundation models. Our approach, NoRA, i.e., LoRA with Nystrom initialization, demonstrates superior performance across various downstream tasks and model scales, from 1B to 7B parameters, in large language and diffusion models.

385. Activation Gradient based Poisoned Sample Detection Against Backdoor Attacks

链接：https://iclr.cc/virtual/2025/poster/29418 abstract： This work studies the task of poisoned sample detection for defending

against data poisoning based backdoor attacks. Its core challenge is finding a generalizable and discriminative metric to distinguish between clean and various types of poisoned samples (e.g., various triggers, various poisoning ratios). Inspired by a common phenomenon in backdoor attacks that the backdoored model tend to map significantly different poisoned and clean samples within the target class to similar activation areas, we introduce a novel perspective of the circular distribution of the gradients w.r.t. sample activation, dubbed gradient circular distribution (GCD). And, we find two interesting observations based on GCD. One is that the GCD of samples in the target class is much more dispersed than that in the clean class. The other is that in the GCD of target class, poisoned and clean samples are clearly separated. Inspired by above two observations, we develop an innovative three-stage poisoned sample detection approach, called Activation Gradient based Poisoned sample Detection (AGPD). First, we calculate GCDs of all classes from the model trained on the untrustworthy dataset. Then, we identify the target class(es) based on the difference on GCD dispersion between target and clean classes. Last, we filter out poisoned samples within the identified target class(es) based on the clear separation between poisoned and clean samples. Extensive experiments under various settings of backdoor attacks demonstrate the superior detection performance of the proposed method to existing poisoned detection approaches according to sample activation-based metrics.

386. Grid Cell-Inspired Fragmentation and Recall for Efficient Map Building

链接：https://iclr.cc/virtual/2025/poster/31475 abstract： Animals and robots navigate through environments by building and refining maps of space. These maps enable functions including navigation back to home, planning, search and foraging. Here, we use observations from neuroscience, specifically the observed fragmentation of grid cell map in compartmentalized spaces, to propose and apply the concept of Fragmentation-and-Recall (FARMap) in the mapping of large spaces. Agents solve the mapping problem by building local maps via a surprisal-based clustering of space, which they use to set subgoals for spatial exploration. Agents build and use a local map to predict their observations; high surprisal leads to a "fragmentation event" that truncates the local map. At these events, the recent local map is placed into long-term memory (LTM) and a different local map is initialized. If observations at a fracture point match observations in one of the stored local maps, that map is recalled (and thus reused) from LTM. The fragmentation points induce a natural online clustering of the larger space, forming a set of intrinsic potential subgoals that are stored in LTM as a topological graph. Agents choose their next subgoal from the set of near and far potential subgoals from within the current local map or LTM, respectively. Thus, local maps guide exploration locally, while LTM promotes global exploration. We demonstrate that FARMap replicates the fragmentation points observed in animal studies. We evaluate FARMap on complex procedurally-generated spatial environments and realistic simulations to demonstrate that this mapping strategy much more rapidly covers the environment (number of agent steps and wall clock time) and is more efficient in active memory usage, without loss of performance.

387. Be More Diverse than the Most Diverse: Optimal Mixtures of Generative Models via Mixture-UCB Bandit Algorithms

链接：https://iclr.cc/virtual/2025/poster/31154 abstract： The availability of multiple training algorithms and architectures for generative models requires a selection mechanism to form a single model over a group of well-trained generation models. The selection task is commonly addressed by identifying the model that maximizes an evaluation score based on the diversity and quality of the generated data. However, such a best-model identification approach overlooks the possibility that a mixture of available models can outperform each individual model. In this work, we numerically show that a mixture of generative models on benchmark image datasets can indeed achieve a better evaluation score (based on FID and KID scores), compared to the individual models. This observation motivates the development of efficient algorithms for selecting the optimal mixture of the models. To address this, we formulate a quadratic optimization problem to find an optimal mixture model achieving the maximum of kernel-based evaluation scores including kernel inception distance (KID) and Rényi kernel entropy (RKE). To identify the optimal mixture of the models using the fewest possible sample queries, we view the selection task as a multi-armed bandit (MAB) problem and propose the Mixture Upper Confidence Bound (Mixture-UCB) algorithm that provably converges to the optimal mixture of the involved models. More broadly, the proposed Mixture-UCB can be extended to optimize every convex quadratic function of the mixture weights in a general MAB setting. We prove a regret bound for the Mixture-UCB algorithm and perform several numerical experiments to show the success of Mixture-UCB in finding the optimal mixture of text and image generative models. The project code is available in the Mixture-UCB Github repository.

388. ReGen: Generative Robot Simulation via Inverse Design

链接：https://iclr.cc/virtual/2025/poster/30399 abstract： Simulation plays a key role in scaling robot learning and validating policies, but constructing simulations remains labor-intensive. In this paper, we introduce ReGen, a generative simulation framework that automates this process using inverse design. Given an agent's behavior (such as a motion trajectory or objective function) and its textual description, we infer the underlying scenarios and environments that could have caused the behavior.Our approach leverages large language models to construct and expand a graph that captures cause-and-effect relationships and relevant entities with properties in the environment, which is then processed to configure a robot simulation environment. Our approach supports (i) augmenting simulations based on ego-agent behaviors, (ii) controllable, counterfactual scenario generation, (iii) reasoning about agent cognition and mental states, and (iv) reasoning with distinct sensing modalities, such as braking due to faulty GPS signals. We demonstrate our method in autonomous driving and robot manipulation tasks, generating more diverse, complex simulated environments compared to existing simulations with high success rates, and enabling controllable generation for corner cases. This approach enhances the validation of robot policies and supports data or simulation augmentation, advancing scalable robot learning for improved generalization and robustness.

389. Self-Supervised Diffusion MRI Denoising via Iterative and Stable Refinement

链接：https://iclr.cc/virtual/2025/poster/27810 abstract： Magnetic Resonance Imaging (MRI), including diffusion MRI (dMRI), serves as a ``microscope'' for anatomical structures and routinely mitigates the influence of low signal-to-noise ratio scans by compromising temporal or spatial resolution. However, these compromises fail to meet clinical demands for both efficiency and precision. Consequently, denoising is a vital preprocessing step, particularly for dMRI, where clean data is unavailable. In this paper, we introduce Di-Fusion, a fully self-supervised denoising method that leverages the latter diffusion steps and an adaptive sampling process. Unlike previous approaches, our single-stage framework achieves efficient and stable training without extra noise model training and offers adaptive and controllable results in the sampling process. Our thorough experiments on real and simulated data demonstrate that Di-Fusion achieves state-of-the-art performance in microstructure modeling, tractography tracking, and other downstream tasks. Code is available at https://github.com/FouierL/Di-Fusion.

390. How Learnable Grids Recover Fine Detail in Low Dimensions: A Neural Tangent Kernel Analysis of Multigrid Parametric Encodings

链接：https://iclr.cc/virtual/2025/poster/30269 abstract： Neural networks that map between low dimensional spaces are ubiquitous incomputer graphics and scientific computing; however, in their naiveimplementation, they are unable to learn high frequency information. We presenta comprehensive analysis comparing the two most common techniques for mitigatingthis spectral bias: Fourier feature encodings (FFE) and multigrid parametricencodings (MPE). FFEs are seen as the standard for low dimensional mappings, butMPEs often outperform them and learn representations with higher resolution andfiner detail. FFE's roots in the Fourier transform, make it susceptible toaliasing if pushed too far, while MPEs, which use a learned grid structure, haveno such limitation. To understand the difference in performance, we use theneural tangent kernel (NTK) to evaluate these encodings through the lens of ananalogous kernel regression. By finding a lower bound on the smallest eigenvalueof the NTK, we prove that MPEs improve a network's performance through thestructure of their grid and not their learnable embedding. This mechanism isfundamentally different from FFEs, which rely solely on their embedding space toimprove performance. Results are empirically validated on a 2D image regressiontask using images taken from 100 synonym sets of ImageNet and 3D implicitsurface regression on objects from the Stanford graphics dataset. Using peaksignal-to-noise ratio (PSNR) and multiscale structural similarity (MS-SSIM) toevaluate how well fine details are learned, we show that the MPE increases theminimum eigenvalue by 8 orders of magnitude over the baseline and 2 orders ofmagnitude over the FFE. The increase in spectrum corresponds to a 15 dB (PSNR) /0.65 (MS-SSIM) increase over baseline and a 12 dB (PSNR) / 0.33 (MS-SSIM) increase over theFFE.

391. RobuRCDet: Enhancing Robustness of Radar-Camera Fusion in Bird's Eye View for 3D Object Detection

链接：https://iclr.cc/virtual/2025/poster/30658 abstract： While recent low-cost radar-camera approaches have shown promising results inmulti-modal 3D object detection, both sensors face challenges from environmen-tal and intrinsic disturbances. Poor lighting or adverse weather conditions de-grade camera performance, while radar suffers from noise and positional ambigu-ity. Achieving robust radar-camera 3D object detection requires consistent perfor-mance across varying conditions, a topic that has not yet been fully explored. Inthis work, we first conduct a systematic analysis of robustness in radar-camera de-tection on five kinds of noises and propose RobuRCDet, a robust object detectionmodel in bird’s eye view (BEV). Specifically, we design a 3D Gaussian Expan-sion (3DGE) module to mitigate inaccuracies in radar points, including position,Radar Cross-Section (RCS), and velocity. The 3DGE uses RCS and velocity priorsto generate a deformable kernel map and variance for kernel size adjustment andvalue distribution. Additionally, we introduce a weather-adaptive fusion module,which adaptively fuses radar and camera features based on camera signal confi-dence. Extensive experiments on the popular benchmark, nuScenes, show thatour RobuRCDet achieves competitive results in regular and noisy conditions. Thesource codes and trained models will be made available.

392. Oryx MLLM: On-Demand Spatial-Temporal Understanding at Arbitrary Resolution

链接：https://iclr.cc/virtual/2025/poster/29834 abstract： Visual data comes in various forms, ranging from small icons of just a few pixels to long videos spanning hours. Existing multi-modal LLMs usually standardize these diverse visual inputs to fixed-resolution images or patches for visual encoders and yield similar numbers of tokens for LLMs. This approach is non-optimal for multimodal understanding and inefficient for processing inputs with long and short visual contents. To solve the problem, we propose Oryx, a unified multimodal architecture for the spatial-temporal understanding of images, videos, and multi-view 3D scenes. Oryx offers an on-demand solution to seamlessly and efficiently process visual inputs with arbitrary spatial sizes and temporal lengths through two core innovations: 1) a pre-trained OryxViT model that can encode images at any resolution into LLM-friendly visual representations; 2) a dynamic compressor module that supports 1x to 16x compression on visual tokens by request. These designs enable Oryx to accommodate extremely long visual contexts, such as videos, with lower resolution and high compression while maintaining high recognition precision for tasks like document understanding with native resolution and no compression. Beyond the architectural improvements, enhanced data curation and specialized training on long-context retrieval and spatial-aware data help Oryx achieve strong capabilities in image, video, and 3D multimodal understanding

simultaneously.

393. AvatarGO: Zero-shot 4D Human-Object Interaction Generation and Animation

链接：https://iclr.cc/virtual/2025/poster/29508 abstract： Recent advancements in diffusion models have led to significant improvements in the generation and animation of 4D full-body human-object interactions (HOI). Nevertheless, existing methods primarily focus on SMPL-based motion generation, which is limited by the scarcity of realistic large-scale interaction data. This constraint affects their ability to create everyday HOI scenes. This paper addresses this challenge using a zero-shot approach with a pre-trained diffusion model. Despite this potential, achieving our goals is difficult due to the diffusion model's lack of understanding of "where" and "how" objects interact with the human body. To tackle these issues, we introduce AvatarGO, a novel framework designed to generate animatable 4D HOI scenes directly from textual inputs. Specifically, 1) for the "where" challenge, we propose LLM-guided contact retargeting, which employs Lang-SAM to identify the contact body part from text prompts, ensuring precise representation of human-object spatial relations. 2) For the "how" challenge, we introduce correspondence-aware motion optimization that constructs motion fields for both human and object models using the linear blend skinning function from SMPL-X. Our framework not only generates coherent compositional motions, but also exhibits greater robustness in handling penetration issues. Extensive experiments with existing methods validate AvatarGO's superior generation and animation capabilities on a variety of human-object pairs and diverse poses. As the first attempt to synthesize 4D avatars with object interactions, we hope AvatarGO could open new doors for human-centric 4D content creation.

394. DynamicCity: Large-Scale 4D Occupancy Generation from Dynamic Scenes

链接：https://iclr.cc/virtual/2025/poster/29953 abstract： Urban scene generation has been developing rapidly recently. However, existing methods primarily focus on generating static and single-frame scenes, overlooking the inherently dynamic nature of real-world driving environments. In this work, we introduce DynamicCity, a novel 4D occupancy generation framework capable of generating large-scale, high-quality dynamic 4D scenes with semantics. DynamicCity mainly consists of two key models. 1) A VAE model for learning HexPlane as the compact 4D representation. Instead of using naive averaging operations, DynamicCity employs a novel Projection Module to effectively compress 4D features into six 2D feature maps for HexPlane construction, which significantly enhances HexPlane fitting quality (up to 12.56 mIoU gain). Furthermore, we utilize an Expansion & Squeeze Strategy to reconstruct 3D feature volumes in parallel, which improves both network training efficiency and reconstruction accuracy than naively querying each 3D point (up to 7.05 mIoU gain, 2.06x training speedup, and 70.84\% memory reduction). 2) A DiT-based diffusion model for HexPlane generation. To make HexPlane feasible for DiT generation, a Padded Rollout Operation is proposed to reorganize all six feature planes of the HexPlane as a squared 2D feature map. In particular, various conditions could be introduced in the diffusion or sampling process, supporting versatile 4D generation applications, such as trajectory- and command-driven generation, inpainting, and layout-conditioned generation. Extensive experiments on the CarlaSC and Waymo datasets demonstrate that DynamicCity significantly outperforms existing state-of-the-art 4D occupancy generation methods across multiple metrics. The code and models have been released to facilitate future research.

395. FasterCache: Training-Free Video Diffusion Model Acceleration with High Quality

链接：https://iclr.cc/virtual/2025/poster/29385 abstract： In this paper, we present \textbf{\textit{FasterCache}}, a novel training-free strategy designed to accelerate the inference of video diffusion models with high-quality generation. By analyzing existing cache-based methods, we observe that \textit{directly reusing adjacent-step features degrades video quality due to the loss of subtle variations}. We further perform a pioneering investigation of the acceleration potential of classifier-free guidance (CFG) and reveal significant redundancy between conditional and unconditional features within the same timestep. Capitalizing on these observations, we introduce FasterCache to substantially accelerate diffusion-based video generation. Our key contributions include a dynamic feature reuse strategy that preserves both feature distinction and temporal continuity, and CFG-Cache which optimizes the reuse of conditional and unconditional outputs to further enhance inference speed without compromising video quality. We empirically evaluate FasterCache on recent video diffusion models. Experimental results show that FasterCache can significantly accelerate video generation (\eg 1.67$\times$ speedup on Vchitect-2.0) while keeping video quality comparable to the baseline, and consistently outperform existing methods in both inference speed and video quality. \textit{Our code will be made public upon publication.}

396. DGQ: Distribution-Aware Group Quantization for Text-to-Image Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29192 abstract： Despite the widespread use of text-to-image diffusion models across various tasks, their computational and memory demands limit practical applications. To mitigate this issue, quantization of diffusion models has been explored. It reduces memory usage and computational costs by compressing weights and activations into lower-bit formats. However, existing methods often struggle to preserve both image quality and text-image alignment, particularly in lower-bit($<$ 8bits) quantization.In this paper, we analyze the challenges associated with quantizing text-to-image

diffusion models from a distributional perspective. Our analysis reveals that activation outliers play a crucial role in determining image quality. Additionally, we identify distinctive patterns in cross-attention scores, which significantly affects text-image alignment.To address these challenges, we propose Distribution-aware Group Quantization (DGQ), a method that identifies and adaptively handles pixel-wise and channel-wise outliers to preserve image quality. Furthermore, DGQ applies prompt-specific logarithmic quantization scales to maintain text-image alignment. Our method demonstrates remarkable performance on datasets such as MS-COCO and PartiPrompts. We are the first to successfully achieve low-bit quantization of text-to-image diffusion models without requiring additional fine-tuning of weight quantization parameters. Code is available at \link{https://github.com/ugonfor/DGQ}.

397. Benchmarking Vision Language Model Unlearning via Fictitious Facial Identity Dataset

链接：https://iclr.cc/virtual/2025/poster/31229 abstract： Machine unlearning has emerged as an effective strategy for forgetting specific information in the training data. However, with the increasing integration of visual data, privacy concerns in Vision Language Models (VLMs) remain underexplored. To address this, we introduce Facial Identity Unlearning Benchmark (FIUBench), a novel VLM unlearning benchmark designed to robustly evaluate the effectiveness of unlearning algorithms under the Right to be Forgotten setting. Specifically, we formulate the VLM unlearning task via constructing the Fictitious Facial Identity VQA dataset and apply a two-stage evaluation pipeline that is designed to precisely control the sources of information and their exposure levels. In terms of evaluation, since VLM supports various forms of ways to ask questions with the same semantic meaning, we also provide robust evaluation metrics including membership inference attacks and carefully designed adversarial privacy attacks to evaluate the performance of algorithms. Through the evaluation of four baseline VLM unlearning algorithms within FIUBench, we find that all methods remain limited in their unlearning performance, with significant trade-offs between model utility and forget quality. Furthermore, our findings also highlight the importance of privacy attacks for robust evaluations. We hope FIUBench will drive progress in developing more effective VLM unlearning algorithms.

398. InverseBench: Benchmarking Plug-and-Play Diffusion Priors for Inverse Problems in Physical Sciences

链接：https://iclr.cc/virtual/2025/poster/29493 abstract： Plug-and-play diffusion priors (PnPDP) have emerged as a promising research direction for solving inverse problems. However, current studies primarily focus on natural image restoration, leaving the performance of these algorithms in scientific inverse problems largely unexplored. To address this gap, we introduce \textsc{InverseBench}, a framework that evaluates diffusion models across five distinct scientific inverse problems. These problems present unique structural challenges that differ from existing benchmarks, arising from critical scientific applications such as optical tomography, medical imaging, black hole imaging, seismology, and fluid dynamics. With \textsc{InverseBench}, we benchmark 14 inverse problem algorithms that use plug-and-play diffusion priors against strong, domain-specific baselines, offering valuable new insights into the strengths and weaknesses of existing algorithms. To facilitate further research and development, we open-source the codebase, along with datasets and pre-trained models, at https://devzhk.github.io/InverseBench/.

399. Lightweight Predictive 3D Gaussian Splats

链接：https://iclr.cc/virtual/2025/poster/29745 abstract： Recent approaches representing 3D objects and scenes using Gaussian splats show increased rendering speed across a variety of platforms and devices. While rendering such representations is indeed extremely efficient, storing and transmitting them is often prohibitively expensive. To represent large-scale scenes, one often needs to store millions of 3D Gaussian, which can occupy up to gigabytes of storage. This creates a significant practical barrier, preventing widespread adoption on resource-constrained devices.In this work, we propose a new representation that dramatically reduces the hard drive footprint while featuring similar or improved quality when compared to the standard 3D Gaussian splats. This representation leverages the inherent feature sharing among splats in the close proximity using a hierarchical tree structure, with which only the parent splats need to be stored. We present a method for constructing tree structures from naturally unstructured point clouds. Additionally, we propose the adaptive tree manipulation to prune the redundant trees in the space, while spawn new ones from the significant children splats during the optimization process. On the benchmark datasets, we achieve 20x storage reduction in hard-drive footprint with improved fidelity compared to the vanilla 3DGS and 2-5x reduction compared to the exiting compact solutions. More importantly, we demonstrate the practical application of our method in real-world rendering on mobile devices and AR glasses.

400. IgGM: A Generative Model for Functional Antibody and Nanobody Design

链接：https://iclr.cc/virtual/2025/poster/27650 abstract： Immunoglobulins are crucial proteins produced by the immune system to identify and bind to foreign substances, playing an essential role in shielding organisms from infections and diseases. Designing specific antibodies opens new pathways for disease treatment. With the rise of deep learning, AI-driven drug design has become possible, leading to several methods for antibody design. However, many of these approaches require additional conditions that differ from real-world scenarios, making it challenging to incorporate them into existing antibody design processes. Here, we introduce IgGM, a generative model for the de novo design of immunoglobulins with functional specificity.

IgGM simultaneously generates antibody sequences and structures for a given antigen, consisting of three core components: a pre-trained language model for extracting sequence features, a feature learning module for identifying pertinent features, and a prediction module that outputs designed antibody sequences and the predicted complete antibody-antigen complex structure. IgGM effectively predicts structures and designs novel antibodies and nanobodies. This makes it highly applicable in a wide range of practical situations related to antibody and nanobody design. Code is available at: https://github.com/TencentAI4S/IgGM.