1401. Bayesian Optimization via Continual Variational Last Layer Training

链接：https://iclr.cc/virtual/2025/poster/31188 abstract： Gaussian Processes (GPs) are widely seen as the state-of-the-art surrogate models for Bayesian optimization (BO) due to their ability to model uncertainty and their performance on tasks where correlations are easily captured (such as those defined by Euclidean metrics) and their ability to be efficiently updated online. However, the performance of GPs depends on the choice of kernel, and kernel selection for complex correlation structures is often difficult or must be made bespoke. While Bayesian neural networks (BNNs) are a promising direction for higher capacity surrogate models, they have so far seen limited use due to poor performance on some problem types. In this paper, we propose an approach which shows competitive performance on many problem types, including some that BNNs typically struggle with. We build on variational Bayesian last layers (VBLLs), and connect training of these models to exact conditioning in GPs. We exploit this connection to develop an efficient online training algorithm that interleaves conditioning and optimization. Our findings suggest that VBLL networks significantly outperform GPs and other BNN architectures on tasks with complex input correlations, and match the performance of well-tuned GPs on established benchmark tasks.

1402. Relation-Aware Diffusion for Heterogeneous Graphs with Partially Observed Features

链接：https://iclr.cc/virtual/2025/poster/29544 abstract： Diffusion-based imputation methods, which impute missing features through the iterative propagation of observed features, have shown impressive performance in homogeneous graphs. However, these methods are not directly applicable to heterogeneous graphs, which have multiple types of nodes and edges, due to two key issues: (1) the presence of nodes with undefined features hinders diffusion-based imputation; (2) treating various edge types equally during diffusion does not fully utilize information contained in heterogeneous graphs. To address these challenges, this paper presents a novel imputation scheme that enables diffusion-based imputation in heterogeneous graphs. Our key idea involves (1) assigning a {\it virtual feature} to an undefined node feature and (2) determining the importance of each edge type during diffusion according to a new criterion. Through experiments, we demonstrate that our virtual feature scheme effectively serves as a bridge between existing diffusion-based methods and heterogeneous graphs, maintaining the advantages of these methods. Furthermore, we confirm that adjusting the importance of each edge type leads to significant performance gains on heterogeneous graphs. Extensive experimental results demonstrate the superiority of our scheme in both semi-supervised node classification and link prediction tasks on heterogeneous graphs with missing rates ranging from low to exceedingly high. The source code is available at https://github.com/daehoum1/hetgfd.

1403. On Rollouts in Model-Based Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29461 abstract： Model-based reinforcement learning (MBRL) seeks to enhance data efficiency by learning a model of the environment and generating synthetic rollouts from it. However, accumulated model errors during these rollouts can distort the data distribution, negatively impacting policy learning and hindering long-term planning. Thus, the accumulation of model errors is a key bottleneck in current MBRL methods. We propose Infoprop, a model-based rollout mechanism that separates aleatoric from epistemic model uncertainty and reduces the influence of the latter on the data distribution. Further, Infoprop keeps track of accumulated model errors along a model rollout and provides termination criteria to limit data corruption. We demonstrate the capabilities of Infoprop in the Infoprop-Dyna algorithm, reporting state-of-the-art performance in Dyna-style MBRL on common MuJoCo benchmark tasks while substantially increasing rollout length and data quality.

1404. TabReD: Analyzing Pitfalls and Filling the Gaps in Tabular Deep Learning Benchmarks

链接：https://iclr.cc/virtual/2025/poster/30020 abstract： Advances in machine learning research drive progress in real-world applications. To ensure this progress, it is important to understand the potential pitfalls on the way from a novel method's success on academic benchmarks to its practical deployment. In this work, we analyze existing tabular deep learning benchmarks and find two common characteristics of tabular data in typical industrial applications that are underrepresented in the datasets usually used for evaluation in the literature.First, in real-world deployment scenarios, distribution of data often changes over time. To account for this distribution drift, time-based train/test splits should be used in evaluation. However, existing academic tabular datasets often lack timestamp metadata to enable such evaluation.Second, a considerable portion of datasets in production settings stem from extensive data acquisition and feature engineering pipelines. This can have an impact on the absolute and relative number of predictive, uninformative, and correlated features compared to academic datasets.In this work, we aim to understand how recent research advances in tabular deep learning transfer to these underrepresented conditions.To this end, we introduce TabReD -- a collection of eight industry-grade tabular datasets. We reassess a large number of tabular ML models and techniques on TabReD. We demonstrate that evaluation on both time-based data splits and richer feature sets leads to different methods ranking, compared to evaluation on random splits and smaller number of features, which are common in academic benchmarks. Furthermore, simple MLP-like architectures and GBDT show the best results on the TabReD datasets, while other methods are less effective in the new setting.

1405. Sports-Traj: A Unified Trajectory Generation Model for Multi-Agent Movement in Sports

链接：https://iclr.cc/virtual/2025/poster/30682 abstract： Understanding multi-agent movement is critical across various fields. The conventional approaches typically focus on separate tasks such as trajectory prediction, imputation, or spatial-temporal recovery. Considering the unique formulation and constraint of each task, most existing methods are tailored for only one, limiting the ability to handle multiple tasks simultaneously, which is a common requirement in real-world scenarios. Another limitation is that widely used public datasets mainly focus on pedestrian movements with casual, loosely connected patterns, where interactions between individuals are not always present, especially at a long distance, making them less representative of more structured environments. To overcome these limitations, we propose a Unified Trajectory Generation model, UniTraj, that processes arbitrary trajectories as masked inputs, adaptable to diverse scenarios in the domain of sports games. Specifically, we introduce a Ghost Spatial Masking (GSM) module, embedded within a Transformer encoder, for spatial feature extraction. We further extend recent State Space Models (SSMs), known as the Mamba model, into a Bidirectional Temporal Mamba (BTM) to better capture temporal dependencies. Additionally, we incorporate a Bidirectional Temporal Scaled (BTS) module to thoroughly scan trajectories while preserving temporal missing relationships. Furthermore, we curate and benchmark three practical sports datasets, Basketball-U, Football-U, and Soccer-U, for evaluation. Extensive experiments demonstrate the superior performance of our model. We hope that our work can advance the understanding of human movement in real-world applications, particularly in sports. Our datasets, code, and model weights are available here https://github.com/colorfulfuture/UniTraj-pytorch.

1406. Scale-Free Graph-Language Models

链接：https://iclr.cc/virtual/2025/poster/28413 abstract： Graph-language models (GLMs) have demonstrated great potential in graph-based semi-supervised learning. A typical GLM consists of two key stages: graph generation and text embedding, which are usually implemented by inferring a latent graph and finetuning a language model (LM), respectively. However, the former often relies on artificial assumptions about the underlying edge distribution, while the latter requires extensive data annotations. To tackle these challenges, this paper introduces a novel GLM that integrates graph generation and text embedding within a unified framework. Specifically, for graph generation, we leverage an inherent characteristic of real edge distribution—the scale-free property—as a structural prior. We unexpectedly find that this natural property can be effectively approximated by a simple k-nearest neighbor (KNN) graph. For text embedding, we develop a graph-based pseudo-labeler that utilizes scale-free graphs to provide complementary supervision for improved LM finetuning. Extensive experiments on representative datasets validate our findings on the scale-free structural approximation of KNN graphs and demonstrate the effectiveness of integrating graph generation and text embedding with a real structural prior. Our code is available at https://github.com/Jianglin954/SFGL.

1407. Accessing Vision Foundation Models via ImageNet-1K

链接：https://iclr.cc/virtual/2025/poster/30012 abstract： Vision foundation models are renowned for the generalization ability due to massive training data. Nevertheless, they demand tremendous training resources, and the training data is often inaccessible, e.g., CLIP, DINOv2, posing great challenges to developing derivatives that could facilitate the research. In this work, we offer a very simple and general solution, named Proteus, to distill foundation models into smaller equivalents on ImageNet-1K without access to the original training data. Specifically, we remove the designs from conventional knowledge distillation settings that result in dataset bias and present three levels of training objectives, i.e., token, patch, and feature, to maximize the efficacy of knowledge transfer. In this manner, Proteus is trained at ImageNet-level costs with surprising ability, facilitating the accessibility of training foundation models for the broader research community. When leveraging DINOv2-g/14 as the teacher, Proteus-L/14 matches the performance of the Oracle method DINOv2-L/14 (142M training data) across 19 benchmarks and outperforms other vision foundation models including CLIP-L/14 (400M), OpenCLIP-L/14 (400M/2B) and SynCLR-L/14 (600M) with a significantly smaller training set of 1.2M images.

1408. Samba: Synchronized Set-of-Sequences Modeling for Multiple Object Tracking

链接：https://iclr.cc/virtual/2025/poster/29814 abstract： Multiple object tracking in complex scenarios - such as coordinated dance performances, team sports, or dynamic animal groups - presents unique challenges. In these settings, objects frequently move in coordinated patterns, occlude each other, and exhibit long-term dependencies in their trajectories. However, it remains a key open research question on how to model long-range dependencies within tracklets, interdependencies among tracklets, and the associated temporal occlusions. To this end, we introduce Samba, a novel linear-time set-of-sequences model designed to jointly process multiple tracklets by synchronizing the multiple selective state-spaces used to model each tracklet. Samba autoregressively predicts the future track query for each sequence while maintaining synchronized long-term memory representations across tracklets. By integrating Samba into a tracking-by-propagation framework, we propose SambaMOTR, the first tracker effectively addressing the aforementioned issues, including long-range dependencies, tracklet interdependencies, and temporal occlusions. Additionally, we introduce an effective technique for dealing with uncertain observations (MaskObs) and an efficient training recipe to scale SambaMOTR to longer sequences. By modeling long-range dependencies and interactions among tracked objects, SambaMOTR implicitly learns to track objects accurately through occlusions without any hand-crafted heuristics. Our approach significantly surpasses prior state-of-the-art on the DanceTrack, BFT, and SportsMOT datasets.

1409. Subgraph Federated Learning for Local Generalization

链接：https://iclr.cc/virtual/2025/poster/29061 abstract： Federated Learning (FL) on graphs enables collaborative model

training to enhance performance without compromising the privacy of each client. However, existing methods often overlook the mutable nature of graph data, which frequently introduces new nodes and leads to shifts in label distribution. Since they focus solely on performing well on each client's local data, they are prone to overfitting to their local distributions (i.e., local overfitting), which hinders their ability to generalize to unseen data with diverse label distributions. In contrast, our proposed method, FedLoG, effectively tackles this issue by mitigating local overfitting. Our model generates global synthetic data by condensing the reliable information from each class representation and its structural information across clients. Using these synthetic data as a training set, we alleviate the local overfitting problem by adaptively generalizing the absent knowledge within each local dataset. This enhances the generalization capabilities of local models, enabling them to handle unseen data effectively. Our model outperforms baselines in our proposed experimental settings, which are designed to measure generalization power to unseen data in practical scenarios. Our code is available at https://github.com/sung-won-kim/FedLoG

1410. Does Refusal Training in LLMs Generalize to the Past Tense?

链接：https://iclr.cc/virtual/2025/poster/29179 abstract： Refusal training is widely used to prevent LLMs from generating harmful, undesirable, or illegal outputs. We reveal a curious generalization gap in the current refusal training approaches: simply reformulating a harmful request in the past tense (e.g., "How to make a Molotov cocktail?" to "How did people make a Molotov cocktail?") is often sufficient to jailbreak many state-of-the-art LLMs. We systematically evaluate this method on Llama-3 8B, Claude-3.5 Sonnet, GPT-3.5 Turbo, Gemma-2 9B, Phi-3-Mini, GPT-4o-mini, GPT-4o, o1-mini, o1-preview, and R2D2 models using GPT-3.5 Turbo as a reformulation model. For example, the success rate of this simple attack on GPT-4o increases from 1\% using direct requests to 88\% using 20 past-tense reformulation attempts on harmful requests from JailbreakBench with GPT-4 as a jailbreak judge. Interestingly, we also find that reformulations in the future tense are less effective, suggesting that refusal guardrails tend to consider past historical questions more benign than hypothetical future questions. Moreover, our experiments on fine-tuning GPT-3.5 Turbo show that defending against past reformulations is feasible when past tense examples are explicitly included in the fine-tuning data. Overall, our findings highlight that the widely used alignment techniques---such as SFT, RLHF, and adversarial training---employed to align the studied models can be brittle and do not always generalize as intended. We provide code and jailbreak artifacts at https://github.com/tml-epfl/llm-past-tense.

1411. Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks

链接：https://iclr.cc/virtual/2025/poster/28755 abstract： We show that even the most recent safety-aligned LLMs are not robust to simple adaptive jailbreaking attacks. First, we demonstrate how to successfully leverage access to logprobs for jailbreaking: we initially design an adversarial prompt template (sometimes adapted to the target LLM), and then we apply random search on a suffix to maximize a target logprob (e.g., of the token ``Sure''), potentially with multiple restarts. In this way, we achieve 100\% attack success rate---according to GPT-4 as a judge---on Vicuna-13B, Mistral-7B, Phi-3-Mini, Nemotron-4-340B, Llama-2-Chat-7B/13B/70B, Llama-3-Instruct-8B, Gemma-7B, GPT-3.5, GPT-4o, and R2D2 from HarmBench that was adversarially trained against the GCG attack. We also show how to jailbreak all Claude models---that do not expose logprobs---via either a transfer or prefilling attack with a 100\% success rate. In addition, we show how to use random search on a restricted set of tokens for finding trojan strings in poisoned models---a task that shares many similarities with jailbreaking---which is the algorithm that brought us the first place in a recent trojan detection competition. The common theme behind these attacks is that adaptivity is crucial: different models are vulnerable to different prompting templates (e.g., R2D2 is very sensitive to in-context learning prompts), some models have unique vulnerabilities based on their APIs (e.g., prefilling for Claude), and in some settings, it is crucial to restrict the token search space based on prior knowledge (e.g., for trojan detection). For reproducibility purposes, we provide the code, logs, and jailbreak artifacts in the JailbreakBench format at https://github.com/tml-epfl/llm-adaptive-attacks.

1412. NNsight and NDIF: Democratizing Access to Open-Weight Foundation Model Internals

链接：https://iclr.cc/virtual/2025/poster/29909 abstract： We introduce NNsight and NDIF, technologies that work in tandem to enable scientific study of the representations and computations learned by very large neural networks. NNsight is an open-source system that extends PyTorch to introduce deferred remote execution. The National Deep Inference Fabric (NDIF) is a scalable inference service that executes NNsight requests, allowing users to share GPU resources and pretrained models. These technologies are enabled by the Intervention Graph, an architecture developed to decouple experimental design from model runtime. Together, this framework provides transparent and efficient access to the internals of deep neural networks such as very large language models (LLMs) without imposing the cost or complexity of hosting customized models individually. We conduct a quantitative survey of the machine learning literature that reveals a growing gap in the study of the internals of large-scale AI. We demonstrate the design and use of our framework to address this gap by enabling a range of research methods on huge models. Finally, we conduct benchmarks to compare performance with previous approaches.Code, documentation, and tutorials are available at https://nnsight.net/.

1413. GETS: Ensemble Temperature Scaling for Calibration in Graph Neural Networks

链接：https://iclr.cc/virtual/2025/poster/28240 abstract： Graph Neural Networks (GNNs) deliver strong classification results but

often suffer from poor calibration performance, leading to overconfidence or underconfidence. This is particularly problematic in high-stakes applications where accurate uncertainty estimates are essential. Existing post-hoc methods, such as temperature scaling, fail to effectively utilize graph structures, while current GNN calibration methods often overlook the potential of leveraging diverse input information and model ensembles jointly. In the paper, we propose Graph Ensemble Temperature Scaling (GETS), a novel calibration framework that combines input and model ensemble strategies within a Graph Mixture-of-Experts (MoE) architecture. GETS integrates diverse inputs, including logits, node features, and degree embeddings, and adaptively selects the most relevant experts for each node's calibration procedure. Our method outperforms state-of-the-art calibration techniques, reducing expected calibration error (ECE) by $\geq$ 25% across 10 GNN benchmark datasets. Additionally, GETS is computationally efficient, scalable, and capable of selecting effective input combinations for improved calibration performance. The implementation is available at https://github.com/ZhuangDingyi/GETS/.

1414. Selective induction Heads: How Transformers Select Causal Structures in Context

链接：https://iclr.cc/virtual/2025/poster/29090 abstract： Transformers have exhibited exceptional capabilities in sequence modelling tasks, leveraging self-attention and in-context learning. Critical to this success are induction heads, attention circuits that enable copying tokens based on their previous occurrences. In this work, we introduce a novel synthetic framework designed to enable the theoretical analysis of transformers' ability to dynamically handle causal structures. Existing works rely on Markov Chains to study the formation of induction heads, revealing how transformers capture causal dependencies and learn transition probabilities in-context. However, they rely on a fixed causal structure that fails to capture the complexity of natural languages, where the relationship between tokens dynamically changes with context. To this end, our framework varies the causal structure through interleaved Markov chains with different lags while keeping the transition probabilities fixed. This setting unveils the formation of Selective Induction Heads, a new circuit that endows transformers with the ability to select the correct causal structure in-context. We empirically demonstrate that attention-only transformers learn this mechanism to predict the next token by identifying the correct lag and copying the corresponding token from the past. We provide a detailed construction of a 3-layer transformer to implement the selective induction head, and a theoretical analysis proving that this mechanism asymptotically converges to the maximum likelihood solution. Our findings advance the theoretical understanding of how transformers select causal structures, providing new insights into their functioning and interpretability.

1415. DreamDistribution: Learning Prompt Distribution for Diverse In-distribution Generation

链接：https://iclr.cc/virtual/2025/poster/28359 abstract： The popularization of Text-to-Image (T2I) diffusion models enables the generation of high-quality images from text descriptions. However, generating diverse customized images with reference visual attributes remains challenging. This work focuses on personalizing T2I diffusion models at a more abstract concept or category level, adapting commonalities from a set of reference images while creating new instances with sufficient variations. We introduce a solution that allows a pretrained T2I diffusion model to learn a set of soft prompts, enabling the generation of novel images by sampling prompts from the learned distribution. These prompts offer text-guided editing capabilities and additional flexibility in controlling variation and mixing between multiple distributions. We also show the adaptability of the learned prompt distribution to other tasks, such as text-to-3D. Finally we demonstrate effectiveness of our approach through quantitative analysis including automatic evaluation and human assessment.

1416. DeepRTL: Bridging Verilog Understanding and Generation with a Unified Representation Model

链接：https://iclr.cc/virtual/2025/poster/31123 abstract： Recent advancements in large language models (LLMs) have shown significant potential for automating hardware description language (HDL) code generation from high-level natural language instructions. While fine-tuning has improved LLMs' performance in hardware design tasks, prior efforts have largely focused on Verilog generation, overlooking the equally critical task of Verilog understanding. Furthermore, existing models suffer from weak alignment between natural language descriptions and Verilog code, hindering the generation of high-quality, synthesizable designs. To address these issues, we present DeepRTL, a unified representation model that excels in both Verilog understanding and generation. Based on CodeT5+, DeepRTL is fine-tuned on a comprehensive dataset that aligns Verilog code with rich, multi-level natural language descriptions. We also introduce the first benchmark for Verilog understanding and take the initiative to apply embedding similarity and GPT Score to evaluate the models' understanding capabilities. These metrics capture semantic similarity more accurately than traditional methods like BLEU and ROUGE, which are limited to surface-level n-gram overlaps. By adapting curriculum learning to train DeepRTL, we enable it to significantly outperform GPT-4 in Verilog understanding tasks, while achieving performance on par with OpenAI's o1-preview model in Verilog generation tasks.

1417. Is In-Context Learning Sufficient for Instruction Following in LLMs?

链接：https://iclr.cc/virtual/2025/poster/29602 abstract： In-context learning (ICL) allows LLMs to learn from examples without changing their weights: this is a particularly promising capability for long-context LLMs that can potentially learn from many examples. Recently, Lin et al. (2024) proposed URIAL, a method using only three in-context examples to align base LLMs,

achieving non-trivial instruction following performance. In this work, we show that, while effective, ICL alignment with URIAL still underperforms compared to instruction fine-tuning on established benchmarks such as MT-Bench and AlpacaEval 2.0 (LC), especially with more capable base LLMs. We then uncover the most relevant elements for successful in-context alignment, finding the crucial role of the decoding parameters. Based on these insights, we show that the approach of URIAL can indeed be improved by adding more, potentially carefully selected, high-quality demonstrations in context, getting closer to the performance of instruct models. Finally, we provide the first, to our knowledge, systematic comparison of ICL and instruction fine-tuning (IFT) for instruction following in the low data regime. Overall, our work advances the understanding of ICL as an alignment technique and its relationship to IFT.

1418. InstantSwap: Fast Customized Concept Swapping across Sharp Shape Differences

链接：https://iclr.cc/virtual/2025/poster/29484 abstract： Recent advances in Customized Concept Swapping (CCS) enable a text-to-image model to swap a concept in the source image with a customized target concept. However, the existing methods still face the challenges of $\textit{\textbf{inconsistency}}$ and $\textit{\textbf{inefficiency}}$. They struggle to maintain consistency in both the foreground and background during concept swapping, especially when the shape difference is large between objects. Additionally, they either require time-consuming training processes or involve redundant calculations during inference. To tackle these issues, we introduce InstantSwap, a new CCS method that aims to handle sharp shape disparity at speed. Specifically, we first extract the bbox of the object in the source image $\textit{automatically}$ based on attention map analysis and leverage the bbox to achieve both foreground and background consistency. For background consistency, we remove the gradient outside the bbox during the swapping process so that the background is free from being modified. For foreground consistency, we employ a cross-attention mechanism to inject semantic information into both source and target concepts inside the box. This helps learn semantic-enhanced representations that encourage the swapping process to focus on the foreground objects. To improve swapping speed, we avoid computing gradients at each timestep but instead calculate them periodically to reduce the number of forward passes, which improves efficiency a lot with a little sacrifice on performance. Finally, we establish a benchmark dataset to facilitate comprehensive evaluation. Extensive evaluations demonstrate the superiority and versatility of InstantSwap.

1419. Swift Hydra: Self-Reinforcing Generative Framework for Anomaly Detection with Multiple Mamba Models

链接：https://iclr.cc/virtual/2025/poster/29771 abstract： Despite a plethora of anomaly detection models developed over the years, their ability to generalize to unseen anomalies remains an issue, particularly in critical systems. This paper aims to address this challenge by introducing Swift Hydra, a new framework for training an anomaly detection method based on generative AI and reinforcement learning (RL). Through featuring an RL policy that operates on the latent variables of a generative model, the framework synthesizes novel and diverse anomaly samples that are capable of bypassing a detection model. These generated synthetic samples are, in turn, used to augment the detection model, further improving its ability to handle challenging anomalies. Swift Hydra also incorporates Mamba models structured as a Mixture of Experts (MoE) to enable scalable adaptation of the number of Mamba experts based on data complexity, effectively capturing diverse feature distributions without increasing the model's inference time. Empirical evaluations on ADBench benchmark demonstrate that Swift Hydra outperforms other state-of-the-art anomaly detection models while maintaining a relatively short inference time. From these results, our research highlights a new and auspicious paradigm of integrating RL and generative AI for advancing anomaly detection.

1420. Learning to Discover Regulatory Elements for Gene Expression Prediction

链接：https://iclr.cc/virtual/2025/poster/29922 abstract： We consider the problem of predicting gene expressions from DNA sequences. A key challenge of this task is to find the regulatory elements that control gene expressions. Here, we introduce Seq2Exp, a Sequence to Expression network explicitly designed to discover and extract regulatory elements that drive target gene expression, enhancing the accuracy of the gene expression prediction. Our approach captures the causal relationship between epigenomic signals, DNA sequences and their associated regulatory elements. Specifically, we propose to decompose the epigenomic signals and the DNA sequence conditioned on the causal active regulatory elements, and apply an information bottleneck with the Beta distribution to combine their effects while filtering out non-causal components. Our experiments demonstrate that Seq2Exp outperforms existing baselines in gene expression prediction tasks and discovers influential regions compared to commonly used statistical methods for peak detection such as MACS3. The source code is released as part of the AIRS library (https://github.com/divelab/AIRS/).

1421. Open-YOLO 3D: Towards Fast and Accurate Open-Vocabulary 3D Instance Segmentation

链接：https://iclr.cc/virtual/2025/poster/30514 abstract： Recent works on open-vocabulary 3D instance segmentation show strong promise but at the cost of slow inference speed and high computation requirements. This high computation cost is typically due to their heavy reliance on aggregated clip features from multi-view, which require computationally expensive 2D

foundation models like Segment Anything (SAM) and CLIP. Consequently, this hampers their applicability in many real-world applications that require both fast and accurate predictions. To this end, we propose a novel open-vocabulary 3D instance segmentation approach, named Open-YOLO 3D, that efficiently leverages only 2D object detection from multi-view RGB images for open-vocabulary 3D instance segmentation. We demonstrate that our proposed Multi-View Prompt Distribution (MVPDist) method makes use of multi-view information to account for misclassification from the object detector to predict a reliable label for 3D instance masks. Furthermore, since projections of 3D object instances are already contained within the 2D bounding boxes, we show that our proposed low granularity label maps, which require only a 2D object detector to construct, are sufficient and very fast to predict prompt IDs for 3D instance masks when used with our proposed MVPDist. We validate our Open-YOLO 3D on two benchmarks, ScanNet200 and Replica, under two scenarios: (i) with ground truth masks, where labels are required for given object proposals, and (ii) with class-agnostic 3D proposals generated from a 3D proposal network. Our Open-YOLO 3D achieves state-of-the-art performance on both datasets while obtaining up to $\sim$16$\times$ speedup compared to the best existing method in literature. On ScanNet200 val. set, our Open-YOLO 3D achieves mean average precision (mAP) of 24.7% while operating at 22 seconds per scene. github.com/aminebdj/OpenYOLO3D

1422. AdaIR: Adaptive All-in-One Image Restoration via Frequency Mining and Modulation

链接：https://iclr.cc/virtual/2025/poster/29955 abstract： In the image acquisition process, various forms of degradation, including noise, blur, haze, and rain, are frequently introduced. These degradations typically arise from the inherent limitations of cameras or unfavorable ambient conditions. To recover clean images from their degraded versions, numerous specialized restoration methods have been developed, each targeting a specific type of degradation. Recently, all-in-one algorithms have garnered significant attention by addressing different types of degradations within a single model without requiring the prior information of the input degradation type. However, most methods purely operate in the spatial domain and do not delve into the distinct frequency variations inherent to different degradation types. To address this gap, we propose an adaptive all-in-one image restoration network based on frequency mining and modulation. Our approach is motivated by the observation that different degradation types impact the image content on different frequency subbands, thereby requiring different treatments for each restoration task. Specifically, we first mine low- and high-frequency information from the input features, guided by the adaptively decoupled spectra of the degraded image. The extracted features are then modulated by a bidirectional operator to facilitate interactions between different frequency components. Finally, the modulated features are merged into the original input for a progressively guided restoration. With this approach, the model achieves adaptive reconstruction by accentuating the informative frequency subbands according to different input degradations. Extensive experiments demonstrate that the proposed method, AdaIR, achieves state-of-the-art performance on different image restoration tasks, including image denoising, dehazing, deraining, motion deblurring, and low-light image enhancement. The code is available at https://github.com/c-yn/AdaIR.

1423. Improved Training Technique for Latent Consistency Models

链接：https://iclr.cc/virtual/2025/poster/29756 abstract： Consistency models are a new family of generative models capable of producing high-quality samples in either a single step or multiple steps. Recently, consistency models have demonstrated impressive performance, achieving results on par with diffusion models in the pixel space. However, the success of scaling consistency training to large-scale datasets, particularly for text-to-image and video generation tasks, is determined by performance in the latent space. In this work, we analyze the statistical differences between pixel and latent spaces, discovering that latent data often contains highly impulsive outliers, which significantly degrade the performance of iCT in the latent space. To address this, we replace Pseudo-Huber losses with Cauchy losses, effectively mitigating the impact of outliers. Additionally, we introduce a diffusion loss at early timesteps and employ optimal transport (OT) coupling to further enhance performance. Lastly, we introduce the adaptive scaling-c scheduler to manage the robust training process and adopt Non-scaling LayerNorm in the architecture to better capture the statistics of the features and reduce outlier impact. With these strategies, we successfully train latent consistency models capable of high-quality sampling with one or two steps, significantly narrowing the performance gap between latent consistency and diffusion models. The implementation is released here: \url{https://github.com/quandao10/sLCT/}

1424. Does Training with Synthetic Data Truly Protect Privacy?

链接：https://iclr.cc/virtual/2025/poster/30530 abstract： As synthetic data becomes increasingly popular in machine learning tasks, numerous methods---without formal differential privacy guarantees---use synthetic data for training. These methods often claim, either explicitly or implicitly, to protect the privacy of the original training data.In this work, we explore four different training paradigms: coreset selection, dataset distillation, data-free knowledge distillation, and synthetic data generated from diffusion models. While all these methods utilize synthetic data for training, they lead to vastly different conclusions regarding privacy preservation. We caution that empirical approaches to preserving data privacy require careful and rigorous evaluation; otherwise, they risk providing a false sense of privacy.

1425. InterMask: 3D Human Interaction Generation via Collaborative Masked Modeling

链接：https://iclr.cc/virtual/2025/poster/29232 abstract： Generating realistic 3D human-human interactions from textual descriptions remains a challenging task. Existing approaches, typically based on diffusion models, often produce results lacking

realism and fidelity. In this work, we introduce *InterMask*, a novel framework for generating human interactions using collaborative masked modeling in discrete space. InterMask first employs a VQ-VAE to transform each motion sequence into a 2D discrete motion token map. Unlike traditional 1D VQ token maps, it better preserves fine-grained spatio-temporal details and promotes *spatial awareness* within each token. Building on this representation, InterMask utilizes a generative masked modeling framework to collaboratively model the tokens of two interacting individuals. This is achieved by employing a transformer architecture specifically designed to capture complex spatio-temporal inter-dependencies. During training, it randomly masks the motion tokens of both individuals and learns to predict them. For inference, starting from fully masked sequences, it progressively fills in the tokens for both individuals. With its enhanced motion representation, dedicated architecture, and effective learning strategy, InterMask achieves state-of-the-art results, producing high-fidelity and diverse human interactions. It outperforms previous methods, achieving an FID of 5.154 (vs 5.535 of in2IN) on the InterHuman dataset and 0.399 (vs 5.207 of InterGen) on the InterX dataset. Additionally, InterMask seamlessly supports reaction generation without the need for model redesign or fine-tuning.

1426. KOR-Bench: Benchmarking Language Models on Knowledge-Orthogonal Reasoning Tasks

链接：https://iclr.cc/virtual/2025/poster/29599 abstract： In this paper, we introduce Knowledge-Orthogonal Reasoning (KOR), a concept aimed at minimizing reliance on domain-specific knowledge, enabling more accurate evaluation of models' reasoning abilities in out-of-distribution settings. Based on this concept, we propose the Knowledge-Orthogonal Reasoning Benchmark (KOR-Bench), encompassing five task categories: Operation, Logic, Cipher, Puzzle, and Counterfactual. KOR-Bench emphasizes models' effectiveness in applying new rule descriptions to solve novel rule-driven questions. O1-Preview and O1-Mini achieve accuracies of 72.88\% and 70.16\%, surpassing Claude-3.5-Sonnet and GPT-4o (58.96\% and 58.00\%), highlighting the effectiveness of KOR-Bench. We perform detailed analyses, identifying bottlenecks in the Cipher task with Stepwise Prompting, where two rounds of Self-Correction yield optimal results. We evaluate performance across three integrated tasks, explore the impact of Tricks on the Puzzle task, and visualize rule-focused attention. Additionally, we conduct an ablation study on dataset size, benchmark correlations, and zero-shot and three-shot "only questions" experiments. KOR-Bench aims to enhance reasoning evaluation and support further research in this area.

1427. MDSGen: Fast and Efficient Masked Diffusion Temporal-Aware Transformers for Open-Domain Sound Generation

链接：https://iclr.cc/virtual/2025/poster/27741 abstract： We introduce MDSGen, a novel framework for vision-guided open-domain sound generation optimized for model parameter size, memory consumption, and inference speed. This framework incorporates two key innovations: (1) a redundant video feature removal module that filters out unnecessary visual information, and (2) a temporal-aware masking strategy that leverages temporal context for enhanced audio generation accuracy. In contrast to existing resource-heavy Unet-based models, MDSGen employs denoising masked diffusion transformers, facilitating efficient generation without reliance on pre-trained diffusion models. Evaluated on the benchmark VGGSound dataset, our smallest model (5M parameters) achieves 97.9% alignment accuracy, using 172x fewer parameters, 371% less memory, and offering 36x faster inference than the current 860M-parameter state-of-the-art model (93.9% accuracy). The larger model (131M parameters) reaches nearly 99% accuracy while requiring 6.5x fewer parameters. These results highlight the scalability and effectiveness of our approach. The code is available at https://bit.ly/mdsgen.

1428. Exact Byte-Level Probabilities from Tokenized Language Models for FIM-Tasks and Model Ensembles

链接：https://iclr.cc/virtual/2025/poster/27677 abstract： Tokenization is associated with many poorly understood shortcomings in language models (LMs), yet remains an important component for long sequence scaling purposes. This work studies how tokenization impacts model performance by analyzing and comparing the stochastic behavior of tokenized models with their byte-level, or token-free, counterparts. We discover that, even when the two models are statistically equivalent, their predictive distributions over the next byte can be substantially different, a phenomenon we term as ``tokenization bias". To fully characterize this phenomenon, we introduce the Byte-Token Representation Lemma, a framework that establishes a mapping between the learned token distribution and its equivalent byte-level distribution. From this result, we develop a next-byte sampling algorithm that eliminates tokenization bias without requiring further training or optimization. In other words, this enables zero-shot conversion of tokenized LMs into statistically equivalent token-free ones. We demonstrate its broad applicability with two use cases: fill-in-the-middle (FIM) tasks and model ensembles. In FIM tasks where input prompts may terminate mid-token, leading to out-of-distribution tokenization, our method mitigates performance degradation and achieves 18\% improvement in FIM coding benchmarks, while consistently outperforming the standard token healing fix. For model ensembles where each model employs a distinct vocabulary, our approach enables seamless integration, resulting in improved performance up to 3.7\% over individual models across various standard baselines in reasoning, knowledge, and coding. Code is available at:https://github.com/facebookresearch/Exact-Byte-Level-Probabilities-from-Tokenized-LMs.

1429. MA^{2}E: Addressing Partial Observability in Multi-Agent Reinforcement Learning with Masked Auto-Encoder

链接：https://iclr.cc/virtual/2025/poster/28562 abstract： Centralized Training and Decentralized Execution (CTDE) is a widely adopted paradigm to solve cooperative multi-agent reinforcement learning (MARL) problems. Despite the successes achieved with CTDE, partial observability still limits cooperation among agents. While previous studies have attempted to overcome this challenge through communication, direct information exchanges could be restricted and introduce additional constraints. Alternatively, if an agent can infer the global information solely from local observations, it can obtain a global view without the need for communication. To this end, we propose the Multi-Agent Masked Auto-Encoder (MA^{2}E), which utilizes the masked auto-encoder architecture to infer the information of other agents from partial observations. By employing masking to learn to reconstruct global information, MA^{2}E serves as an inference module for individual agents within the CTDE framework. MA^{2}E can be easily integrated into existing MARL algorithms and has been experimentally proven to be effective across a wide range of environments and algorithms.

1430. Multi-Task Corrupted Prediction for Learning Robust Audio-Visual Speech Representation

链接：https://iclr.cc/virtual/2025/poster/29375 abstract： Audio-visual speech recognition (AVSR) incorporates auditory and visual modalities to improve recognition accuracy, particularly in noisy environments where audio-only speech systems are insufficient. While previous research has largely addressed audio disruptions, few studies have dealt with visual corruptions, e.g., lip occlusions or blurred videos, which are also detrimental. To address this real-world challenge, we propose CAV2vec, a novel self-supervised speech representation learning framework particularly designed to handle audio-visual joint corruption. CAV2vec employs a self-distillation approach with a corrupted prediction task, where the student model learns to predict clean targets, generated by the teacher model, with corrupted input frames. Specifically, we suggest a unimodal multi-task learning, which distills cross-modal knowledge and aligns the corrupted modalities, by predicting clean audio targets with corrupted videos, and clean video targets with corrupted audios. This strategy mitigates the dispersion in the representation space caused by corrupted modalities, leading to more reliable and robust audio-visual fusion. Our experiments on robust AVSR benchmarks demonstrate that the corrupted representation learning method significantly enhances recognition accuracy across generalized environments involving various types of corruption. Our code is available at https://github.com/sungnyun/cav2vec.

1431. Multi-Perspective Data Augmentation for Few-shot Object Detection

链接：https://iclr.cc/virtual/2025/poster/28258 abstract： Recent few-shot object detection (FSOD) methods have focused on augmenting synthetic samples for novel classes, show promising results to the rise of diffusion models. However, the diversity of such datasets is often limited in representativeness because they lack awareness of typical and hard samples, especially in the context of foreground and background relationships. To tackle this issue, we propose a Multi-Perspective Data Augmentation (MPAD) framework. In terms of foreground-foreground relationships, we propose in-context learning for object synthesis (ICOS) with bounding box adjustments to enhance the detail and spatial information of synthetic samples. Inspired by the large margin principle, support samples play a vital role in defining class boundaries. Therefore, we design a Harmonic Prompt Aggregation Scheduler (HPAS) to mix prompt embeddings at each time step of the generation process in diffusion models, producing hard novel samples. For foreground-background relationships, we introduce a Background Proposal method (BAP) to sample typical and hard backgrounds. Extensive experiments on multiple FSOD benchmarks demonstrate the effectiveness of our approach. Our framework significantly outperforms traditional methods, achieving an average increase of 17.5% in nAP50 over the baseline on PASCAL VOC.

1432. Automated Filtering of Human Feedback Data for Aligning Text-to-Image Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/30748 abstract： Fine-tuning text-to-image diffusion models with human feedback is an effective method for aligning model behavior with human intentions. However, this alignment process often suffers from slow convergence due to the large size and noise present in human feedback datasets. In this work, we propose FiFA, a novel automated data filtering algorithm designed to enhance the fine-tuning of diffusion models using human feedback datasets with direct preference optimization (DPO). Specifically, our approach selects data by solving an optimization problem to maximize three components: preference margin, text quality, and text diversity. The concept of preference margin is used to identify samples that are highly informative in addressing the noisy nature of feedback dataset, which is calculated using a proxy reward model. Additionally, we incorporate text quality, assessed by large language models to prevent harmful contents, and consider text diversity through a k-nearest neighbor entropy estimator to improve generalization. Finally, we integrate all these components into an optimization process, with approximating the solution by assigning importance score to each data pair and selecting the most important ones. As a result, our method efficiently filters data automatically, without the need for manual intervention, and can be applied to any large-scale dataset. Experimental results show that FiFA significantly enhances training stability and achieves better performance, being preferred by humans 17% more, while using less than 0.5% of the full data and thus 1% of the GPU hours compared to utilizing full human feedback datasets.

1433. TOP-ERL: Transformer-based Off-Policy Episodic Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29902 abstract： This work introduces Transformer-based Off-Policy Episodic

Reinforcement Learning (TOP-ERL), a novel algorithm that enables off-policy updates in the ERL framework. In ERL, policies predict entire action trajectories over multiple time steps instead of single actions at every time step. These trajectories are typically parameterized by trajectory generators such as Movement Primitives (MP), allowing for smooth and efficient exploration over long horizons while capturing high-level temporal correlations. However, ERL methods are often constrained to on-policy frameworks due to the difficulty of evaluating state-action values for entire action sequences, limiting their sample efficiency and preventing the use of more efficient off-policy architectures. TOP-ERL addresses this shortcoming by segmenting long action sequences and estimating the state-action values for each segment using a transformer-based critic architecture alongside an n-step return estimation. These contributions result in efficient and stable training that is reflected in the empirical results conducted on sophisticated robot learning environments. TOP-ERL significantly outperforms state-of-the-art RL methods. Thorough ablation studies additionally show the impact of key design choices on the model performance.

1434. Rapid Selection and Ordering of In-Context Demonstrations via Prompt Embedding Clustering

链接：https://iclr.cc/virtual/2025/poster/31211 abstract： While Large Language Models (LLMs) excel at in-context learning (ICL) using just a few demonstrations, their performances are sensitive to demonstration orders. The reasons behind this sensitivity remain poorly understood. In this paper, we investigate the prompt embedding space to bridge the gap between the order sensitivity of ICL with inner workings of decoder-only LLMs, uncovering the clustering property: prompts sharing the first and last demonstrations have closer embeddings, with first-demonstration clustering usually being stronger in practice. We explain this property through extensive theoretical analyses and empirical evidences. Our finding suggests that the positional encoding and the causal attention mask are key contributors to the clustering phenomenon. Leveraging this clustering insight, we introduce Cluster-based Search, a novel method that accelerates the selection and ordering of demonstrations in self-adaptive ICL settings. Our approach substantially decreases the time complexity from factorial to quadratic, saving 92% to nearly 100% execution time while maintaining comparable performance to exhaustive search.

1435. Plug, Play, and Generalize: Length Extrapolation with Pointer-Augmented Neural Memory

链接：https://iclr.cc/virtual/2025/poster/31462 abstract： We introduce Pointer-Augmented Neural Memory (PANM), a versatile module designed to enhance neural networks' ability to process symbols and extend their capabilities to longer data sequences. PANM integrates an external neural memory utilizing novel physical addresses and pointer manipulation techniques, emulating human and computer-like symbol processing abilities. PANM facilitates operations like pointer assignment, dereferencing, and arithmetic by explicitly employing physical pointers for memory access. This module can be trained end-to-end on sequence data, empowering various sequential models, from simple recurrent networks to large language models (LLMs). Our experiments showcase PANM's exceptional length extrapolation capabilities and its enhancement of recurrent neural networks in symbol processing tasks, including algorithmic reasoning and Dyck language recognition. PANM enables Transformers to achieve up to 100% generalization accuracy in compositional learning tasks and significantly improves performance in mathematical reasoning, question answering, and machine translation. Notably, the generalization effectiveness scales with stronger backbone models, as evidenced by substantial performance gains when we test LLMs finetuned with PANM for tasks up to 10-100 times longer than the training data.

1436. FlickerFusion: Intra-trajectory Domain Generalizing Multi-agent Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29932 abstract： Multi-agent reinforcement learning has demonstrated significant potential in addressing complex cooperative tasks across various real-world applications. However, existing MARL approaches often rely on the restrictive assumption that the number of entities (e.g., agents, obstacles) remains constant between training and inference. This overlooks scenarios where entities are dynamically removed or $\textit{added}$ $\textit{during}$ the inference trajectory—a common occurrence in real-world environments like search and rescue missions and dynamic combat situations. In this paper, we tackle the challenge of intra-trajectory dynamic entity composition under zero-shot out-of-domain (OOD) generalization, where such dynamic changes cannot be anticipated beforehand. Our empirical studies reveal that existing MARL methods suffer $\textit{significant}$ performance degradation and increased uncertainty in these scenarios. In response, we propose FlickerFusion, a novel OOD generalization method that acts as a $\textit{universally}$ applicable augmentation technique for MARL backbone methods. FlickerFusion stochastically drops out parts of the observation space, emulating being in-domain when inferenced OOD. The results show that FlickerFusion not only achieves superior inference rewards but also $\textit{uniquely}$ reduces uncertainty vis-à-vis the backbone, compared to existing methods. Benchmarks, implementations, and model weights are organized and open-sourced at $\texttt{\href{flickerfusion305.github.io}{\textbf{flickerfusion305.github.io}}}$, accompanied by ample demo video renderings.

1437. EDiT: A Local-SGD-Based Efficient Distributed Training Method for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/27758 abstract： Distributed training methods are crucial for large language models (LLMs). However, existing distributed training methods often suffer from communication bottlenecks, stragglers, and limited

elasticity, particularly in heterogeneous or large-scale environments. Local SGD methods have been proposed to address these issues, but their effectiveness remains limited to small-scale training due to additional memory overhead and lack of concerns on efficiency and stability. To tackle these issues, we propose EDiT, an innovative Efficient Distributed Training method that combines a tailored Local SGD approach with model sharding techniques to enhance large-scale training efficiency. EDiT performs layer-wise parameter synchronization during forward pass, reducing communication and memory overhead and enabling overlap. Besides, EDiT employs a pseudo gradient penalty strategy to suppress loss spikes, which ensures training stability and improves performance. Additionally, we introduce A-EDiT, a fully asynchronous variant of EDiT that accommodates heterogeneous clusters. Building on EDiT/A-EDiT, we conduct a series of experiments to validate large-scale asynchronous training for LLMs, accompanied by comprehensive analyses. Experimental results demonstrate the superior performance of EDiT/A-EDiT, establishing them as robust solutions for distributed LLM training in diverse computational ecosystems. The code is available at Atorch codebase: https://github.com/intelligent-machine-learning/atorch/tree/main/atorch/local_sgd.

1438. Deconstructing Denoising Diffusion Models for Self-Supervised Learning

链接：https://iclr.cc/virtual/2025/poster/30665 abstract： In this study, we examine the representation learning abilities of Denoising Diffusion Models (DDM) that were originally purposed for image generation. Our philosophy is to deconstruct a DDM, gradually transforming it into a classical Denoising Autoencoder (DAE). This deconstructive process allows us to explore how various components of modern DDMs influence self-supervised representation learning. We observe that only a very few modern components are critical for learning good representations, while many others are nonessential. Our study ultimately arrives at an approach that is highly simplified and to a large extent resembles a classical DAE. We hope our study will rekindle interest in a family of classical methods within the realm of modern self-supervised learning.

1439. An Image is Worth More Than 16x16 Patches: Exploring Transformers on Individual Pixels

链接：https://iclr.cc/virtual/2025/poster/28029 abstract： This work does not introduce a new method. Instead, we present an interesting finding that questions the necessity of the inductive bias of locality in modern computer vision architectures. Concretely, we find that vanilla Transformers can operate by directly treating each individual pixel as a token and achieve highly performant results. This is substantially different from the popular design in Vision Transformer, which maintains the inductive bias from ConvNets towards local neighborhoods (e.g., by treating each 16x16 patch as a token). We showcase the effectiveness of pixels-as-tokens across three well-studied computer vision tasks: supervised learning for classification and regression, self-supervised learning via masked autoencoding, and image generation with diffusion models. Although it's computationally less practical to directly operate on individual pixels, we believe the community must be made aware of this surprising piece of knowledge when devising the next generation of neural network architectures for computer vision.

1440. Student-Informed Teacher Training

链接：https://iclr.cc/virtual/2025/poster/30431 abstract： Imitation learning with a privileged teacher has proven effective for learning complex control behaviors from high-dimensional inputs, such as images. In this framework, a teacher is trained with privileged task information, while a student tries to predict the actions of the teacher with more limited observations, e.g., in a robot navigation task, the teacher might have access to distances to nearby obstacles, while the student only receives visual observations of the scene. However, privileged imitation learning faces a key challenge: the student might be unable to imitate the teacher's behavior due to partial observability. This problem arises because the teacher is trained without considering if the student is capable of imitating the learned behavior. To address this teacher-student asymmetry, we propose a framework for joint training of the teacher and student policies, encouraging the teacher to learn behaviors that can be imitated by the student despite the latters' limited access to information and its partial observability. Based on the performance bound in imitation learning, we add (i) the approximated action difference between teacher and student as a penalty term to the reward function of the teacher, and (ii) a supervised teacher-student alignment step. We motivate our method with a maze navigation task and demonstrate its effectiveness on complex vision-based quadrotor flight and manipulation tasks.

1441. Stochastic variance-reduced Gaussian variational inference on the Bures-Wasserstein manifold

链接：https://iclr.cc/virtual/2025/poster/28700 abstract： Optimization in the Bures-Wasserstein space has been gaining popularity in the machine learning community since it draws connections between variational inference and Wasserstein gradient flows. The variational inference objective function of Kullback–Leibler divergence can be written as the sum of the negative entropy and the potential energy, making forward-backward Euler the method of choice. Notably, the backward step admits a closed-form solution in this case, facilitating the practicality of the scheme. However, the forward step is not exact since the Bures-Wasserstein gradient of the potential energy involves "intractable" expectations. Recent approaches propose using the Monte Carlo method -- in practice a single-sample estimator -- to approximate these terms, resulting in high variance and poor performance. We propose a novel variance-reduced estimator based on the principle of control variates. We theoretically show that this estimator has a smaller variance than the Monte-Carlo estimator in scenarios of interest. We also prove that variance reduction helps improve the optimization bounds of the current analysis. We demonstrate that the proposed estimator

gains order-of-magnitude improvements over the previous Bures-Wasserstein methods.

1442. Generalization through variance: how noise shapes inductive biases in diffusion models

链接：https://iclr.cc/virtual/2025/poster/30806 abstract： How diffusion models generalize beyond their training set is not known, and is somewhat mysterious given two facts: the optimum of the denoising score matching (DSM) objective usually used to train diffusion models is the score function of the training distribution; and the networks usually used to learn the score function are expressive enough to learn this score to high accuracy. We claim that a certain feature of the DSM objective—the fact that its target is not the training distribution's score, but a noisy quantity only equal to it in expectation—strongly impacts whether and to what extent diffusion models generalize. In this paper, we develop a mathematical theory that partly explains this 'generalization through variance' phenomenon. Our theoretical analysis exploits a physics-inspired path integral approach to compute the distributions typically learned by a few paradigmatic under- and overparameterized diffusion models. We find that the distributions diffusion models effectively learn to sample from resemble their training distributions, but with `gaps' filled in, and that this inductive bias is due to the covariance structure of the noisy target used during training. We also characterize how this inductive bias interacts with feature-related inductive biases.

1443. A deep inverse-mapping model for a flapping robotic wing

链接：https://iclr.cc/virtual/2025/poster/31165 abstract： In systems control, the dynamics of a system are governed by modulating its inputs to achieve a desired outcome. For example, to control the thrust of a quad-copter propeller the controller modulates its rotation rate, relying on a straightforward mapping between the input rotation rate and the resulting thrust. This mapping can be inverted to determine the rotation rate needed to generate a desired thrust. However, in complex systems, such as flapping-wing robots where intricate fluid motions are involved, mapping inputs (wing kinematics) to outcomes (aerodynamic forces) is nontrivial and inverting this mapping for real-time control is computationally impractical. Here, we report a machine-learning solution for the inverse mapping of a flapping-wing system based on data from an experimental system we have developed. Our model learns the input wing motion required to generate a desired aerodynamic force outcome. We used a sequence-to-sequence model tailored for time-series data and augmented it with a novel adaptive-spectrum layer that implements representation learning in the frequency domain. To train our model, we developed a flapping wing system that simultaneously measures the wing's aerodynamic force and its 3D motion using high-speed cameras. We demonstrate the performance of our system on an additional open-source dataset of a flapping wing in a different flow regime. Results show superior performance compared with more complex state-of-the-art transformer-based models, with 11\% improvement on the test datasets median loss. Moreover, our model shows superior inference time, making it practical for onboard robotic control. Our open-source data and framework may improve modeling and real-time control of systems governed by complex dynamics, from biomimetic robots to biomedical devices.

1444. Capability Localization: Capabilities Can be Localized rather than Individual Knowledge

链接：https://iclr.cc/virtual/2025/poster/28895 abstract： Large scale language models have achieved superior performance in tasks related to natural language processing, however, it is still unclear how model parameters affect performance improvement. Previous studies assumed that individual knowledge is stored in local parameters, and the storage form of individual knowledge is dispersed parameters, parameter layers, or parameter chains, which are not unified. We found through fidelity and reliability evaluation experiments that individual knowledge cannot be localized. Afterwards, we constructed a dataset for decoupling experiments and discovered the potential for localizing data commonalities. To further reveal this phenomenon, this paper proposes a Commonality Neuron Localization (CNL) method, which successfully locates commonality neurons and achieves a neuron overlap rate of 96.42% on the GSM8K dataset. Finally, we have demonstrated through cross data experiments that commonality neurons are a collection of capability neurons that possess the capability to enhance performance. Our code is available at https://github.com/nlpkeg/Capability-Neuron-Localization.

1445. Learning Successor Features with Distributed Hebbian Temporal Memory

链接：https://iclr.cc/virtual/2025/poster/27836 abstract： This paper presents a novel approach to address the challenge of online sequence learning for decision making under uncertainty in non-stationary, partially observable environments. The proposed algorithm, Distributed Hebbian Temporal Memory (DHTM), is based on the factor graph formalism and a multi-component neuron model. DHTM aims to capture sequential data relationships and make cumulative predictions about future observations, forming Successor Features (SFs). Inspired by neurophysiological models of the neocortex, the algorithm uses distributed representations, sparse transition matrices, and local Hebbian-like learning rules to overcome the instability and slow learning of traditional temporal memory algorithms such as RNN and HMM. Experimental results show that DHTM outperforms LSTM, RWKV and a biologically inspired HMM-like algorithm, CSCG, on non-stationary data sets. Our results suggest that DHTM is a promising approach to address the challenges of online sequence learning and planning in dynamic environments.

1446. Dist Loss: Enhancing Regression in Few-Shot Region through

Distribution Distance Constraint

链接：https://iclr.cc/virtual/2025/poster/29253 abstract： Imbalanced data distributions are prevalent in real-world scenarios, presenting significant challenges in both classification and regression tasks. This imbalance often causes deep learning models to overfit in regions with abundant data (manyshot regions) while underperforming in regions with sparse data (few-shot regions). Such characteristics limit the applicability of deep learning models across various domains, notably in healthcare, where rare cases often carry greater clinical significance. While recent studies have highlighted the benefits of incorporating distributional information in imbalanced classification tasks, similar strategies have been largely unexplored in imbalanced regression. To address this gap, we propose Dist Loss, a novel loss function that integrates distributional information into model training by jointly optimizing the distribution distance between model predictions and target labels, alongside sample-wise prediction errors. This dual-objective approach encourages the model to balance its predictions across different label regions, leading to significant improvements in accuracy in fewshot regions. We conduct extensive experiments across three datasets spanning computer vision and healthcare: IMDB-WIKI-DIR, AgeDB-DIR, and ECG-KDIR. The results demonstrate that Dist Loss effectively mitigates the impact of imbalanced data distributions, achieving state-of-the-art performance in few-shot regions. Furthermore, Dist Loss is easy to integrate and complements existing methods. To facilitate further research, we provide our implementation at https://github.com/Ngk03/DIR-Dist-Loss.

1447. JudgeLM: Fine-tuned Large Language Models are Scalable Judges

链接：https://iclr.cc/virtual/2025/poster/27760 abstract： Evaluating Large Language Models (LLMs) in open-ended scenarios is challenging because existing benchmarks and metrics can not measure them comprehensively. To address this problem, we propose to fine-tune LLMs as scalable judges (JudgeLM) to evaluate LLMs efficiently and effectively in open-ended benchmarks. We first propose a comprehensive, large-scale, high-quality dataset containing task seeds, LLMs-generated answers, and GPT-4-generated judgments for fine-tuning high-performance judges, as well as a new benchmark for evaluating the judges. We train JudgeLM at different scales from 7B, 13B, to 33B parameters, and conduct a systematic analysis of its capabilities and behaviors. We then analyze the key biases in fine-tuning LLM as a judge and consider them as position bias, knowledge bias, and format bias. To address these issues, JudgeLM introduces a bag of techniques including swap augmentation, reference support, and reference drop, which clearly enhance the judge's performance. JudgeLM obtains the state-of-the-art judge performance on both the existing PandaLM benchmark and our proposed new benchmark. Our JudgeLM is efficient and the JudgeLM-7B only needs 3 minutes to judge 5K samples with 8 A100 GPUs. JudgeLM obtains high agreement with the teacher judge, achieving an agreement exceeding 90% that even surpasses human-to-human agreement. JudgeLM also demonstrates extended capabilities in being judges of the single answer, multimodal models, multiple answers, multi-turn chat, etc.

1448. Robust Transfer of Safety-Constrained Reinforcement Learning Agents

链接：https://iclr.cc/virtual/2025/poster/28164 abstract： Reinforcement learning (RL) often relies on trial and error, which may cause undesirable outcomes. As a result, standard RL is inappropriate for safety-critical applications. To address this issue, one may train a safe agent in a controlled environment (where safety violations are allowed) and then transfer it to the real world (where safety violations may have disastrous consequences). Prior work has made this transfer safe as long as the new environment preserves the safety-related dynamics. However, in most practical applications, differences or shifts in dynamics between the two environments are inevitable, potentially leading to safety violations after the transfer. This work aims to guarantee safety even when the new environment has different (safety-related) dynamics. In other words, we aim to make the process of safe transfer robust. Our methodology (1) robustifies an agent in the controlled environment and (2) provably provides---under mild assumption---a safe transfer to new environments. The empirical evaluation shows that this method yields policies that are robust against changes in dynamics, demonstrating safety after transfer to a new environment.

1449. On Discriminative Probabilistic Modeling for Self-Supervised Representation Learning

链接：https://iclr.cc/virtual/2025/poster/28153 abstract： We study the discriminative probabilistic modeling on a continuous domain for the data prediction task of (multimodal) self-supervised representation learning. To address the challenge of computing the integral in the partition function for each anchor data, we leverage the multiple importance sampling (MIS) technique for robust Monte Carlo integration, which can recover InfoNCE-based contrastive loss as a special case. Within this probabilistic modeling framework, we conduct generalization error analysis to reveal the limitation of current InfoNCE-based contrastive loss for self-supervised representation learning and derive insights for developing better approaches by reducing the error of Monte Carlo integration. To this end, we propose a novel non-parametric method for approximating the sum of conditional probability densities required by MIS through convex optimization, yielding a new contrastive objective for self-supervised representation learning. Moreover, we design an efficient algorithm for solving the proposed objective. We empirically compare our algorithm to representative baselines on the contrastive image-language pretraining task. Experimental results on the CC3M and CC12M datasets demonstrate the superior overall performance of our algorithm. Our code is available at https://github.com/bokun-wang/NUCLR.

1450. SplineGS: Learning Smooth Trajectories in Gaussian Splatting for

Dynamic Scene Reconstruction

链接：https://iclr.cc/virtual/2025/poster/28059 abstract： Reconstructing complex scenes with deforming objects for novel view synthesis is a challenging task. Recent works have addressed this with 3D Gaussian Splatting, which effectively reconstructs static scenes with high quality in short training time, by adding specialized modules for the deformations of Gaussian blobs. However, designing an effective deformation module that incorporates appropriate spatiotemporal inductive biases still remains unresolved. To address this issue, we propose SplineGS in this paper, which utilizes non-uniform rational B-splines (NURBS), an extension of B-spline, to represent temporally smooth deformation. A set of representative trajectories are learned based on NURBS, and the individual trajectories of Gaussian blobs are represented as linear combinations of these trajectories for spatial smoothness. The weights of the combinations are trained based on a multi-resolution hash table and an MLP, with the positions of the Gaussian blobs as the keys. Thanks to this design, the proposed method does not need any regularizers for trajectories, which enables efficient training. Experiments demonstrate that the proposed method provides competitive performance over the existing methods with much shorter training time.

1451. SLMRec: Distilling Large Language Models into Small for Sequential Recommendation

链接：https://iclr.cc/virtual/2025/poster/30304 abstract： Sequential Recommendation (SR) task involves predicting the next item a user is likely to interact with, given their past interactions. The SR models examine the sequence of a user's actions to discern more complex behavioral patterns and temporal dynamics. Recent research demonstrates the great impact of LLMs on sequential recommendation systems, either viewing sequential recommendation as language modeling or serving as the backbone for user representation. Although these methods deliver outstanding performance, there is scant evidence of the necessity of a large language model and how large the language model is needed, especially in the sequential recommendation scene. Meanwhile, due to the huge size of LLMs, it is inefficient and impractical to apply a LLM-based model in real-world platforms that often need to process billions of traffic logs daily. In this paper, we explore the influence of LLMs' depth by conducting extensive experiments on large-scale industry datasets. Surprisingly, our motivational experiments reveal that most intermediate layers of LLMs are redundant, indicating that pruning the remaining layers can still maintain strong performance.Motivated by this insight, we empower small language models for SR, namely SLMRec, which adopt a simple yet effective knowledge distillation method. Moreover, SLMRec is orthogonal to other post-training efficiency techniques, such as quantization and pruning, so that they can be leveraged in combination. Comprehensive experimental results illustrate that the proposed SLMRec model attains the best performance using only 13\% of the parameters found in LLM-based recommendation models while simultaneously achieving up to 6.6x and 8.0x speedups in training and inference time costs, respectively. Besides, we provide a theoretical justification for why small language models can perform comparably to large language models in SR.

1452. PADRe: A Unifying Polynomial Attention Drop-in Replacement for Efficient Vision Transformer

链接：https://iclr.cc/virtual/2025/poster/29269 abstract： We present Polynomial Attention Drop-in Replacement (PADRe), a novel and unifying framework designed to replace the conventional self-attention mechanism in transformer models. Notably, several recent alternative attention mechanisms, including Hyena, Mamba, SimA, Conv2Former, and Castling-ViT, can be viewed as specific instances of our PADRe framework. PADRe leverages polynomial functions and draws upon established results from approximation theory, enhancing computational efficiency without compromising accuracy. PADRe's key components include multiplicative nonlinearities, which we implement using straightforward, hardware-friendly operations such as Hadamard products, incurring only linear computational and memory costs. PADRe further avoids the need for using complex functions such as Softmax, yet it maintains comparable or superior accuracy compared to traditional self-attention. We assess the effectiveness of PADRe as a drop-in replacement for self-attention across diverse computer vision tasks. These tasks include image classification, image-based 2D object detection, and 3D point cloud object detection. Empirical results demonstrate that PADRe runs significantly faster than the conventional self-attention (11x~43x faster on server GPU and mobile NPU) while maintaining similar accuracy when substituting self-attention in the transformer models.

1453. Accelerated Over-Relaxation Heavy-Ball Method: Achieving Global Accelerated Convergence with Broad Generalization

链接：https://iclr.cc/virtual/2025/poster/29597 abstract： The heavy-ball momentum method accelerates gradient descent with a momentum term but lacks accelerated convergence for general smooth strongly convex problems. This work introduces the Accelerated Over-Relaxation Heavy-Ball (AOR-HB) method, the first variant with provable global and accelerated convergence for such problems. AOR-HB closes a long-standing theoretical gap, extends to composite convex optimization and min-max problems, and achieves optimal complexity bounds. It offers three key advantages: (1) broad generalization ability, (2) potential to reshape acceleration techniques, and (3) conceptual clarity and elegance compared to existing methods.

1454. BaB-ND: Long-Horizon Motion Planning with Branch-and-Bound and Neural Dynamics

链接：https://iclr.cc/virtual/2025/poster/30104 abstract： Neural-network-based dynamics models learned from observational data have shown strong predictive capabilities for scene dynamics in robotic manipulation tasks. However, their inherent non-linearity presents significant challenges for effective planning. Current planning methods, often dependent on extensive sampling or local gradient descent, struggle with long-horizon motion planning tasks involving complex contact events.In this paper, we present a GPU-accelerated branch-and-bound (BaB) framework for motion planning in manipulation tasks that require trajectory optimization over neural dynamics models. Our approach employs a specialized branching heuristic to divide the search space into sub-domains and applies a modified bound propagation method, inspired by the state-of-the-art neural network verifier α,β-CROWN, to efficiently estimate objective bounds within these sub-domains. The branching process guides planning effectively, while the bounding process strategically reduces the search space.Our framework achieves superior planning performance, generating high-quality state-action trajectories and surpassing existing methods in challenging, contact-rich manipulation tasks such as non-prehensile planar pushing with obstacles, object sorting, and rope routing in both simulated and real-world settings. Furthermore, our framework supports various neural network architectures, ranging from simple multilayer perceptrons to advanced graph neural dynamics models, and scales efficiently with different model sizes.

1455. Towards Generalizable Reinforcement Learning via Causality-Guided Self-Adaptive Representations

链接：https://iclr.cc/virtual/2025/poster/29113 abstract： General intelligence requires quick adaptation across tasks. While existing reinforcement learning (RL) methods have made progress in generalization, they typically assume only distribution changes between source and target domains. In this paper, we explore a wider range of scenarios where not only the distribution but also the environment spaces may change. For example, in the CoinRun environment, we train agents from easy levels and generalize them to difficulty levels where there could be new enemies that have never occurred before. To address this challenging setting, we introduce a causality-guided self-adaptive representation-based approach, called CSR, that equips the agent to generalize effectively across tasks with evolving dynamics. Specifically, we employ causal representation learning to characterize the latent causal variables within the RL system. Such compact causal representations uncover the structural relationships among variables, enabling the agent to autonomously determine whether changes in the environment stem from distribution shifts or variations in space, and to precisely locate these changes. We then devise a three-step strategy to fine-tune the causal model under different scenarios accordingly. Empirical experiments show that CSR efficiently adapts to the target domains with only a few samples and outperforms state-of-the-art baselines on a wide range of scenarios, including our simulated environments, CartPole, CoinRun and Atari games.

1456. PolyhedronNet: Representation Learning for Polyhedra with Surface-attributed Graph

链接：https://iclr.cc/virtual/2025/poster/30542 abstract： Ubiquitous geometric objects can be precisely and efficiently represented as polyhedra. The transformation of a polyhedron into a vector, known as polyhedra representation learning, is crucial for manipulating these shapes with mathematical and statistical tools for tasks like classification, clustering, and generation. Recent years have witnessed significant strides in this domain, yet most efforts focus on the vertex sequence of a polyhedron, neglecting the complex surface modeling crucial in real-world polyhedral objects.This study proposes \textbf{PolyhedronNet}, a general framework tailored for learning representations of 3D polyhedral objects. We propose the concept of the surface-attributed graph to seamlessly model the vertices, edges, faces, and their geometric interrelationships within a polyhedron. To effectively learn the representation of the entire surface-attributed graph, we first propose to break it down into local rigid representations to effectively learn each local region's relative positions against the remaining regions without geometric information loss. Subsequently, we propose PolyhedronGNN to hierarchically aggregate the local rigid representation via intra-face and inter-face geometric message passing modules, to obtain a global representation that minimizes information loss while maintaining rotation and translation invariance.Our experimental evaluations on four distinct datasets, encompassing both classification and retrieval tasks, substantiate PolyhedronNet's efficacy in capturing comprehensive and informative representations of 3D polyhedral objects.

1457. Gaussian Splatting Lucas-Kanade

链接：https://iclr.cc/virtual/2025/poster/28968 abstract： Gaussian Splatting and its dynamic extensions are effective for reconstructing 3D scenes from 2D images when there is significant camera movement to facilitate motion parallax and when scene objects remain relatively static. However, in many real-world scenarios, these conditions are not met. As a consequence, data-driven semantic and geometric priors have been favored as regularizers, despite their bias toward training data and their neglect of broader movement dynamics.Departing from this practice, we propose a novel analytical approach that adapts the classical Lucas-Kanade method to dynamic Gaussian splatting. By leveraging the intrinsic properties of the forward warp field network, we derive an analytical velocity field that, through time integration, facilitates accurate scene flow computation. This enables the precise enforcement of motion constraints on warp fields, thus constraining both 2D motion and 3D positions of the Gaussians. Our method excels in reconstructing highly dynamic scenes with minimal camera movement, as demonstrated through experiments on both synthetic and real-world scenes.

1458. Test of Time: A Benchmark for Evaluating LLMs on Temporal Reasoning

链接：https://iclr.cc/virtual/2025/poster/31040 abstract： Large language models (LLMs) have showcased remarkable reasoning capabilities, yet they remain susceptible to errors, particularly in temporal reasoning tasks involving complex temporal logic. Existing research has explored LLM performance on temporal reasoning using diverse datasets and benchmarks. However, these studies often rely on real-world data that LLMs may have encountered during pre-training or employ anonymization techniques that can inadvertently introduce factual inconsistencies. In this work, we address these limitations by introducing novel synthetic datasets specifically designed to assess LLM temporal reasoning abilities in various scenarios. The diversity of question types across these datasets enables systematic investigation into the impact of the problem structure, size, question type, fact order, and other factors on LLM performance. Our findings provide valuable insights into the strengths and weaknesses of current LLMs in temporal reasoning tasks. To foster further research in this area, we will open-source the datasets and evaluation framework used in our experiments.

1459. Surgical, Cheap, and Flexible: Mitigating False Refusal in Language Models via Single Vector Ablation

链接：https://iclr.cc/virtual/2025/poster/29621 abstract： Training a language model to be both helpful and harmless requires careful calibration of refusal behaviours: Models should refuse to follow malicious instructions or give harmful advice (e.g."how do I kill someone?"), but they should not refuse safe requests, even if they superficially resemble unsafe ones (e.g. "how do I kill a Python process?"). Avoiding such false refusal, as prior work has shown, is challenging even for highly-capable language models. In this paper, we propose a simple and surgical method for mitigating false refusal in language models via single vector ablation. For a given model, we extract a false refusal vector and show that ablating this vector reduces false refusal rate while preserving the model's safety and general capabilities. We also show that our approach can be used for fine-grained calibration of model safety. Our approach is training-free and model-agnostic, making it useful for mitigating the problem of false refusal in current and future language models.

1460. Mind the GAP: Glimpse-based Active Perception improves generalization and sample efficiency of visual reasoning

链接：https://iclr.cc/virtual/2025/poster/28691 abstract： Human capabilities in understanding visual relations are far superior to those of AI systems, especially for previously unseen objects. For example, while AI systems struggle to determine whether two such objects are visually the same or different, humans can do so with ease. Active vision theories postulate that the learning of visual relations is grounded in actions that we take to fixate objects and their parts by moving our eyes. In particular, the low-dimensional spatial information about the corresponding eye movements is hypothesized to facilitate the representation of relations between different image parts. Inspired by these theories, we develop a system equipped with a novel Glimpse-based Active Perception (GAP) that sequentially glimpses at the most salient regions of the input image and processes them at high resolution. Importantly, our system leverages the locations stemming from the glimpsing actions, along with the visual content around them, to represent relations between different parts of the image. The results suggest that the GAP is essential for extracting visual relations that go beyond the immediate visual content. Our approach reaches state-of-the-art performance on several visual reasoning tasks being more sample-efficient, and generalizing better to out-of-distribution visual inputs than prior models.

1461. SCBench: A KV Cache-Centric Analysis of Long-Context Methods

链接：https://iclr.cc/virtual/2025/poster/28803 abstract： Long-context Large Language Models (LLMs) have enabled numerous downstream applications but also introduced significant challenges related to computational and memory efficiency. To address these challenges, optimizations for long-context inference have been developed, centered around the KV cache. However, existing benchmarks often evaluate in single-request, neglecting the full lifecycle of the KV cache in real-world use. This oversight is particularly critical, as KV cache reuse has become widely adopted in LLMs inference frameworks, such as vLLM and SGLang, as well as by LLM providers, including OpenAI, Microsoft, Google, and Anthropic. To address this gap, we introduce SCBENCH (SharedContextBENCH), a comprehensive benchmark for evaluating long-context methods from a KV cache centric perspective: 1) KV cache generation, 2) KV cache compression, 3) KV cache retrieval, and 4) KV cache loading. Specifically, SCBench uses test examples with shared context, ranging 12 tasks with two shared context modes, covering four categories of long-context capabilities: string retrieval, semantic retrieval, global information, and multi-task. With SCBench, we provide an extensive KV cache-centric analysis of eight categories long-context solutions, including Gated Linear RNNs (Codestal-Mamba), Mamba-Attention hybrids (Jamba-1.5-Mini), and efficient methods such as sparse attention, KV cache dropping, quantization, retrieval, loading, and prompt compression. The evaluation is conducted on six Transformer-based long-context LLMs: Llama-3.1-8B/70B, Qwen2.5-72B/32B, Llama-3-8B-262K, and GLM-4-9B. Our findings show that sub-O(n) memory methods suffer in multi-turn scenarios, while sparse encoding with O(n) memory and sub-O(n^2) pre-filling computation perform robustly. Dynamic sparsity yields more expressive KV caches than static patterns, and layer-level sparsity in hybrid architectures reduces memory usage with strong performance. Additionally, we identify attention distribution shift issues in long-generation scenarios.

1462. Exponential Topology-enabled Scalable Communication in Multi-agent Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/30519 abstract： In cooperative multi-agent reinforcement learning (MARL), well-designed communication protocols can effectively facilitate consensus among agents, thereby enhancing task performance. Moreover, in large-scale multi-agent systems commonly found in real-world applications, effective communication plays an even more critical role due to the escalated challenge of partial observability compared to smaller-scale setups. In this work, we endeavor to develop a scalable communication protocol for MARL. Unlike previous methods that focus on selecting optimal pairwise communication links—a task that becomes increasingly complex as the number of agents grows—we adopt a global perspective on communication topology design. Specifically, we propose utilizing the exponential topology to enable rapid information dissemination among agents by leveraging its small-diameter and small-size properties. This approach leads to a scalable communication protocol, named ExpoComm. To fully unlock the potential of exponential graphs as communication topologies, we employ memory-based message processors and auxiliary tasks to ground messages, ensuring that they reflect global information and benefit decision-making. Extensive experiments on large-scale cooperative benchmarks, including MAgent and Infrastructure Management Planning, demonstrate the superior performance and robust zero-shot transferability of ExpoComm compared to existing communication strategies. Thecode is publicly available at https://github.com/LXXXXR/ExpoComm.

1463. On the self-verification limitations of large language models on reasoning and planning tasks

链接：https://iclr.cc/virtual/2025/poster/31014 abstract： There has been considerable divergence of opinion on the reasoning abilities of Large Language Models (LLMs).While the initial optimism that reasoning might emerge automatically with scale has been tempered thanks to a slew of counterexamples--ranging from multiplication to simple planning--there persists a wide spread belief that LLMs can self-critique and improve their own solutions in an iterative fashion.This belief seemingly rests on the assumption that verification of correctness should be easier than generation--a rather classical argument from computational complexity--which should be irrelevant to LLMs to the extent that what they are doing is approximate retrieval.In this paper, we set out to systematically investigate the effectiveness of iterative prompting in the context of reasoning and planning.We present a principled empirical study of the performance of GPT-4 in three domains: Game of 24, Graph Coloring, and STRIPS planning.We experiment both with the model critiquing its own answers and with an external correct reasoner verifying proposed solutions.In each case, we analyze whether the content of criticisms actually affects bottom line performance, and whether we can ablate elements of the augmented system without losing performance. We observe significant performance collapsewith self-critique and significant performance gains with sound external verification.We also note that merely re-prompting with a sound verifier maintains most of the benefits of more involved setups.

1464. Multi-modal brain encoding models for multi-modal stimuli

链接：https://iclr.cc/virtual/2025/poster/31247 abstract： Despite participants engaging in unimodal stimuli, such as watching images or silent videos, recent work has demonstrated that multi-modal Transformer models can predict visual brain activity impressively well, even with incongruent modality representations. This raises the question of how accurately these multi-modal models can predict brain activity when participants are engaged in multi-modal stimuli. As these models grow increasingly popular, their use in studying neural activity provides insights into how our brains respond to such multi-modal naturalistic stimuli, i.e., where it separates and integrates information across modalities through a hierarchy of early sensory regions to higher cognition (language regions). We investigate this question by using multiple unimodal and two types of multi-modal models—cross-modal and jointly pretrained—to determine which type of models is more relevant to fMRI brain activity when participants are engaged in watching movies (videos with audio). We observe that both types of multi-modal models show improved alignment in several language and visual regions. This study also helps in identifying which brain regions process unimodal versus multi-modal information. We further investigate the contribution of each modality to multi-modal alignment by carefully removing unimodal features one by one from multi-modal representations, and find that there is additional information beyond the unimodal embeddings that is processed in the visual and language regions. Based on this investigation, we find that while for cross-modal models, their brain alignment is partially attributed to the video modality; for jointly pretrained models, it is partially attributed to both the video and audio modalities. These findings serve as strong motivation for the neuro-science community to investigate the interpretability of these models for deepening our understanding of multi-modal information processing in brain.

1465. Brain Bandit: A Biologically Grounded Neural Network for Efficient Control of Exploration

链接：https://iclr.cc/virtual/2025/poster/29647 abstract： How to balance between exploration and exploitation in an uncertain environment is a central challenge in reinforcement learning. In contrast, humans and animals have demonstrated superior exploration efficiency in novel environments. To understand how the brain’s neural network controls exploration under uncertainty, we analyzed the dynamical systems model of a biological neural network that controls explore-exploit decisions during foraging. Mathematically, this model (named the Brain Bandit Net, or BBN) is a special type of stochastic continuous Hopfield network. We show through theory and simulation that BBN can perform posterior sampling of action values with a tunable bias towards or against uncertain options. We then demonstrate that, in multi-armed bandit (MAB) tasks, BBN can generate probabilistic choice behavior with a flexible uncertainty bias resembling human and animal choice patterns. In addition to its high efficiency in MAB tasks, BBN can also be embedded with reinforcement learning algorithms to accelerate learning in MDP tasks. Altogether, our findings reveal the theoretical foundation for efficient exploration in biological neural networks and propose a general, brain-inspired algorithm for enhancing exploration in RL.

1466. Robustness Auditing for Linear Regression: To Singularity and Beyond

链接：https://iclr.cc/virtual/2025/poster/29431 abstract： It has recently been discovered that the conclusions of many highly influential econometrics studies can be overturned by removing a very small fraction of their samples (often less than 0.5%). These conclusions are typically based on the results of one or more Ordinary Least Squares (OLS) regressions, raising the question: given a dataset, can we certify the robustness of an OLS fit on this dataset to the removal of a given number of samples?Brute-force techniques quickly break down even on small datasets. Existing approaches which go beyond brute force either can only find candidate small subsets to remove (but cannot certify their non-existence) [BGM20, KZC21], are computationally intractable beyond low dimensional settings [MR22], or require very strong assumptions on the data distribution and too many samples to give reasonable bounds in practice [BP21, FH23]. We present an efficient algorithm for certifying the robustness of linear regressions to removals of samples. We implement our algorithm and run it on several landmark econometrics datasets with hundreds of dimensions and tens of thousands of samples, giving the first non-trivial certificates of robustness to sample removal for datasets of dimension 4 or greater. We prove that under distributional assumptions on a dataset, the bounds produced by our algorithm are tight up to a $1 + o(1)$ multiplicative factor.

1467. Time-MoE: Billion-Scale Time Series Foundation Models with Mixture of Experts

链接：https://iclr.cc/virtual/2025/poster/28954 abstract： Deep learning for time series forecasting has seen significant advancements over the past decades. However, despite the success of large-scale pre-training in language and vision domains, pre-trained time series models remain limited in scale and operate at a high cost, hindering the development of larger capable forecasting models in real-world applications. In response, we introduce Time-MoE, a scalable and unified architecture designed to pre-train larger, more capable forecasting foundation models while reducing inference costs. By leveraging a sparse mixture-of-experts (MoE) design, Time-MoE enhances computational efficiency by activating only a subset of networks for each prediction, reducing computational load while maintaining high model capacity. This allows Time-MoE to scale effectively without a corresponding increase in inference costs. Time-MoE comprises a family of decoder-only transformer models that operate in an auto-regressive manner and support flexible forecasting horizons with varying input context lengths. We pre-trained these models on our newly introduced large-scale data Time-300B, which spans over 9 domains and encompassing over 300 billion time points. For the first time, we scaled a time series foundation model up to 2.4 billion parameters, achieving significantly improved forecasting precision. Our results validate the applicability of scaling laws for training tokens and model size in the context of time series forecasting. Compared to dense models with the same number of activated parameters or equivalent computation budgets, our models consistently outperform them by large margin. These advancements position Time-MoE as a state-of-the-art solution for tackling real-world time series forecasting challenges with superior capability, efficiency, and flexibility. Code is available at https://github.com/Time-MoE/Time-MoE

1468. Joint Graph Rewiring and Feature Denoising via Spectral Resonance

链接：https://iclr.cc/virtual/2025/poster/27686 abstract： When learning from graph data, the graph and the node features both give noisy information about the node labels. In this paper we propose an algorithm to jointly denoise the features and rewire the graph (JDR), which improves the performance of downstream node classification graph neural nets (GNNs). JDR works by aligning the leading spectral spaces of graph and feature matrices. It approximately solves the associated non-convex optimization problem in a way that handles graphs with multiple classes and different levels of homophily or heterophily. We theoretically justify JDR in a stylized setting and show that it consistently outperforms existing rewiring methods on a wide range of synthetic and real-world node classification tasks.

1469. Expected Sliced Transport Plans

链接：https://iclr.cc/virtual/2025/poster/29773 abstract： The optimal transport (OT) problem has gained significant traction in modern machine learning for its ability to: (1) provide versatile metrics, such as Wasserstein distances and their variants, and (2) determine optimal couplings between probability measures. To reduce the computational complexity of OT solvers, methods like entropic regularization and sliced optimal transport have been proposed. The sliced OT framework improves efficiency by comparing one-dimensional projections (slices) of high-dimensional distributions. However, despite their computational efficiency, sliced-Wasserstein approaches lack a transportation plan between the input measures, limiting their use in scenarios requiring explicit coupling. In this paper, we address two key questions: Can a transportation plan be constructed between two probability measures using the sliced transport framework? If so, can this plan be used to define a metric between the measures? We propose a ‘lifting’ operation to extend one-dimensional optimal transport plans back to the original space of the measures. By computing the expectation of these lifted plans, we derive a new transportation plan, termed expected sliced transport (EST) plans. We further prove that using the EST plan to weight the sum of the individual Euclidean costs $\|x - y\|^p$ for moving from x to y results in a valid metric between the input discrete probability measures. Finally, we demonstrate the connection between our approach and the recently proposed min-SWGG, along with illustrative numerical examples that support our theoretical findings.

1470. Cross the Gap: Exposing the Intra-modal Misalignment in CLIP via Modality Inversion

链接：https://iclr.cc/virtual/2025/poster/29411 abstract： Pre-trained multi-modal Vision-Language Models like CLIP are widely used off-the-shelf for a variety of applications. In this paper, we show that the common practice of individually exploiting the text or image encoders of these powerful multi-modal models is highly suboptimal for intra-modal tasks like image-to-image retrieval. We argue that this is inherently due to the CLIP-style inter-modal contrastive loss that does not enforce any intra-modal constraints, leading to what we call intra-modal misalignment. To demonstrate this, we leverage two optimization-based modality inversion techniques that map representations from their input modality to the complementary one without any need for auxiliary data or additional trained adapters. We empirically show that, in the intra-modal tasks of image-to-image and text-to-text retrieval, approaching these tasks inter-modally significantly improves performance with respect to intra-modal baselines on more than fifteen datasets. Additionally, we demonstrate that approaching a native inter-modal task (e.g. zero-shot image classification) intra-modally decreases performance, further validating our findings. Finally, we show that incorporating an intra-modal term in the pre-training objective or narrowing the modality gap between the text and image feature embedding spaces helps reduce the intra-modal misalignment. The code is publicly available at: https://github.com/miccunifi/Cross-the-Gap.

1471. HAMSTER: Hierarchical Action Models for Open-World Robot Manipulation

链接：https://iclr.cc/virtual/2025/poster/28776 abstract： Large foundation models have shown strong open-world generalization to complex problems in vision and language, but similar levels of generalization have yet to be achieved in robotics. One fundamental challenge is the lack of robotic data, which are typically obtained through expensive on-robot operation. A promising remedy is to leverage cheaper, off-domain data such as action-free videos, hand-drawn sketches, or simulation data. In this work, we posit that hierarchical vision-language-action (VLA) models can be more effective in utilizing off-domain data than standard monolithic VLA models that directly finetune vision-language models (VLMs) to predict actions.In particular, we study a class of hierarchical VLA models, where the high-level VLM is finetuned to produce a coarse 2D path indicating the desired robot end-effector trajectory given an RGB image and a task description. The intermediate 2D path prediction is then served as guidance to the low-level, 3D-aware control policy capable of precise manipulation. Doing so alleviates the high-level VLM from fine-grained action prediction, while reducing the low-level policy's burden on complex task-level reasoning.We show that, with the hierarchical design, the high-level VLM can transfer across significant domain gaps between the off-domain finetuning data and real-robot testing scenarios, including differences in embodiments, dynamics, visual appearances, and task semantics, etc.In the real-robot experiments, we observe an average of 20% improvement in success rate across seven different axes of generalization over OpenVLA, representing a 50% relative gain.Visual results are provided at: https://hamster-robot.github.io/

1472. Regularization by Texts for Latent Diffusion Inverse Solvers

链接：https://iclr.cc/virtual/2025/poster/29505 abstract：

1473. Partial Gromov-Wasserstein Metric

链接：https://iclr.cc/virtual/2025/poster/28140 abstract： The Gromov-Wasserstein (GW) distance has gained increasing interest in the machine learning community in recent years, as it allows for the comparison of measures in different metric spaces. To overcome the limitations imposed by the equal mass requirements of the classical GW problem, researchers have begun exploring its application in unbalanced settings. However, Unbalanced GW (UGW) can only be regarded as a discrepancy rather than a rigorous metric/distance between two metric measure spaces (mm-spaces). In this paper, we propose a particular case of the UGW problem, termed Partial Gromov-Wasserstein (PGW). We establish that PGW is a well-defined metric between mm-spaces and discuss its theoretical properties, including the existence of a minimizer for the PGW problem and the relationship between PGW and GW, among others. We then propose two variants of the Frank-Wolfe algorithm for solving the PGW problem and show that they are mathematically and computationally equivalent. Moreover, based on our PGW metric, we introduce the analogous concept of barycenters for mm-spaces. Finally, we validate the effectiveness of our PGW metric and related solvers in applications such as shape matching, shape retrieval, and shape interpolation, comparing them against existing baselines. Our code is available at https://github.com/mint-vu/PGW_Metric.

1474. CFG++: Manifold-constrained Classifier Free Guidance for Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/30421 abstract： Classifier-free guidance (CFG) is a fundamental tool in modern diffusion models for text-guided generation. Although effective, CFG has notable drawbacks. For instance, DDIM with CFG lacks invertibility, complicating image editing; furthermore, high guidance scales, essential for high-quality outputs, frequently result in issues like mode collapse. Contrary to the widespread belief that these are inherent limitations of diffusion models, this paper reveals that the problems actually stem from the off-manifold phenomenon associated with CFG, rather than the diffusion models themselves. More specifically, inspired by the recent advancements of diffusion model-based inverse problem solvers (DIS), we reformulate text-guidance as an inverse problem with a text-conditioned score matching loss and develop CFG++, a

novel approach that tackles the off-manifold challenges inherent in traditional CFG. CFG++ features a surprisingly simple fix to CFG, yet it offers significant improvements, including better sample quality for text-to-image generation, invertibility, smaller guidance scales, reduced etc. Furthermore, CFG++ enables seamless interpolation between unconditional and conditional sampling at lower guidance scales, consistently outperforming traditional CFG at all scales. Moreover, CFG++ can be easily integrated into the high-order diffusion solvers and naturally extends to distilled diffusion models. Experimental results confirm that our method significantly enhances performance in text-to-image generation, DDIM inversion, editing, and solving inverse problems, suggesting a wide-ranging impact and potential applications in various fields that utilize text guidance. Project Page: https://cfgpp-diffusion.github.io/anon

1475. Improving Large Language Model Planning with Action Sequence Similarity

链接：https://iclr.cc/virtual/2025/poster/28022 abstract： Planning is essential for artificial intelligence systems to look ahead and proactively determine a course of actions to reach objectives in the virtual and real world. Recent work on large language models (LLMs) sheds light on their planning capability in various tasks. However, it remains unclear what signals in the context influence the model performance. In this work, we explore how to improve the model planning capability through in-context learning (ICL), specifically, what signals can help select the exemplars. Through extensive experiments, we observe that commonly used problem similarity may result in false positives with drastically different plans, which can mislead the model. In response, we propose to sample and filter exemplars leveraging plan side action sequence similarity (AS). We propose GRASE-DC: a two-stage pipeline that first re-samples high AS exemplars and then curates the selected exemplars with dynamic clustering on AS to achieve a balance of relevance and diversity. Our experimental result confirms that GRASE-DC achieves significant performance improvement on various planning tasks (up to ~11-40 point absolute accuracy improvement with 27.3% fewer exemplars needed on average). With GRASE-DC* + VAL, where we iteratively apply GRASE-DC with a validator, we are able to even boost the performance by 18.9% more.Extensive analysis validates the consistent performance improvement of GRASE-DC with various backbone LLMs and on both classical planning and natural language planning benchmarks. GRASE-DC can further boost the planning accuracy by ~24 absolute points on harder problems using simpler problems as exemplars over a random baseline. This demonstrates its ability to generalize to out-of-distribution problems.

1476. Differential Transformer

链接：https://iclr.cc/virtual/2025/poster/29792 abstract： Transformer tends to overallocate attention to irrelevant context. In this work, we introduce Diff Transformer, which amplifies attention to the relevant context while canceling noise. Specifically, the differential attention mechanism calculates attention scores as the difference between two separate softmax attention maps. The subtraction cancels noise, promoting the emergence of sparse attention patterns. Experimental results on language modeling show that Diff Transformer outperforms Transformer in various settings of scaling up model size and training tokens. More intriguingly, it offers notable advantages in practical applications, such as long-context modeling, key information retrieval, hallucination mitigation, in-context learning, and reduction of activation outliers. By being less distracted by irrelevant context, Diff Transformer can mitigate hallucination in question answering and text summarization. For in-context learning, Diff Transformer not only enhances accuracy but is also more robust to order permutation, which was considered as a chronic robustness issue. The results position Diff Transformer as a highly effective and promising architecture for large language models.

1477. Bayesian Analysis of Combinatorial Gaussian Process Bandits

链接：https://iclr.cc/virtual/2025/poster/30976 abstract： We consider the combinatorial volatile Gaussian process (GP) semi-bandit problem. Each round, an agent is provided a set of available base arms and must select a subset of them to maximize the long-term cumulative reward. We study the Bayesian setting and provide novel Bayesian cumulative regret bounds for three GP-based algorithms: GP-UCB, GP-BayesUCB and GP-TS. Our bounds extend previous results for GP-UCB and GP-TS to the \emph{infinite}, \emph{volatile} and \emph{combinatorial} setting, and to the best of our knowledge, we provide the first regret bound for GP-BayesUCB. Volatile arms encompass other widely considered bandit problems such as contextual bandits.Furthermore, we employ our framework to address the challenging real-world problem of online energy-efficient navigation, where we demonstrate its effectiveness compared to the alternatives.

1478. Atomas: Hierarchical Adaptive Alignment on Molecule-Text for Unified Molecule Understanding and Generation

链接：https://iclr.cc/virtual/2025/poster/28439 abstract： Molecule-and-text cross-modal representation learning has emerged as a promising direction for enhancing the quality of molecular representation, thereby improving performance in various scientific fields. However, most approaches employ a global alignment approach to learn the knowledge from different modalities that may fail to capture fine-grained information, such as molecule-and-text fragments and stereoisomeric nuances, which is crucial for downstream tasks. Furthermore, it is incapable of modeling such information using a similar global alignment strategy due to the lack of annotations about the fine-grained fragments in the existing dataset.In this paper, we propose Atomas, a hierarchical molecular representation learning framework that jointly learns representations from SMILES strings and text. We design a Hierarchical Adaptive Alignment model to automatically learn the fine-grained fragment correspondence between two modalities and align these representations at three semantic levels. Atomas's end-to-end training framework supports understanding and generating molecules, enabling a wider range of downstream tasks. Atomas achieves superior

performance across 12 tasks on 11 datasets, outperforming 11 baseline models thus highlighting the effectiveness and versatility of our method. Scaling experiments further demonstrate Atomas's robustness and scalability. Moreover, visualization and qualitative analysis, validated by human experts, confirm the chemical relevance of our approach. Codes are released on ~\url{https://github.com/yikunpku/Atomas}.

1479. Demystifying the Token Dynamics of Deep Selective State Space Models

链接：https://iclr.cc/virtual/2025/poster/28232 abstract： Selective state space models (SSM), such as Mamba, have gained prominence for their effectiveness in modeling sequential data. Despite their outstanding empirical performance, a comprehensive theoretical understanding of deep selective SSM remains elusive, hindering their further development and adoption for applications that need high fidelity. In this paper, we investigate the dynamical properties of tokens in a pre-trained Mamba model. In particular, we derive the dynamical system governing the continuous-time limit of the Mamba model and characterize the asymptotic behavior of its solutions. In the one-dimensional case, we prove that only one of the following two scenarios happens: either all tokens converge to zero, or all tokens diverge to infinity. We provide criteria based on model parameters to determine when each scenario occurs. For the convergent scenario, we empirically verify that this scenario negatively impacts the model's performance. For the divergent scenario, we prove that different tokens will diverge to infinity at different rates, thereby contributing unequally to the updates during model training. Based on these investigations, we propose two refinements for the model: excluding the convergent scenario and reordering tokens based on their importance scores, both aimed at improving practical performance. Our experimental results validate these refinements, offering insights into enhancing Mamba's effectiveness in real-world applications.

1480. Visual Haystacks: A Vision-Centric Needle-In-A-Haystack Benchmark

链接：https://iclr.cc/virtual/2025/poster/30702 abstract： Large Multimodal Models (LMMs) have made significant strides in visual question-answering for single images. Recent advancements like long-context LMMs have allowed them to ingest larger, or even multiple, images. However, the ability to process a large number of visual tokens does not guarantee effective retrieval and reasoning for multi-image question answering (MIQA), especially in real-world applications like photo album searches or satellite imagery analysis. In this work, we first assess the limitations of current benchmarks for long-context LMMs. We address these limitations by introducing a new vision-centric, long-context benchmark, "Visual Haystacks (VHs)". We comprehensively evaluate both open-source and proprietary models on VHs, and demonstrate that these models struggle when reasoning across potentially unrelated images, perform poorly on cross-image reasoning, as well as exhibit biases based on the placement of key information within the context window. Towards a solution, we introduce MIRAGE (Multi-Image Retrieval Augmented Generation), an open-source, lightweight visual-RAG framework that processes up to 10k images on a single 40G A100 GPU—far surpassing the 1k-image limit of contemporary models. MIRAGE demonstrates up to 13% performance improvement over existing open-source LMMs on VHs, sets a new state-of-the-art on the RetVQA multi-image QA benchmark, and achieves competitive performance on single-image QA with state-of-the-art LMMs. Our dataset, model, and code are available at: https://visual-haystacks.github.io.

1481. Language-Image Models with 3D Understanding

链接：https://iclr.cc/virtual/2025/poster/27715 abstract： Multi-modal large language models (MLLMs) have shown incredible capabilities in a variety of 2D vision and language tasks. We extend MLLMs' perceptual capabilities to ground and reason about images in 3-dimensional space. To that end, we first develop a large-scale pretraining dataset for 2D and 3D called LV3D by combining multiple existing 2D and 3D recognition datasets under a common task formulation: as multi-turn question-answering. Next, we introduce a new MLLM named CUBE-LLM and pre-train it on LV3D. We show that pure data scaling makes a strong 3D perception capability without 3D specific architectural design or training objective. CUBE-LLM exhibits intriguing properties similar to LLMs: (1) CUBE-LLM can apply chain-of-thought prompting to improve 3D understanding from 2D context information. (2) CUBE-LLM can follow complex and diverse instructions and adapt to versatile input and output formats. (3) CUBE-LLM can be visually prompted such as 2D box or a set of candidate 3D boxes from specialists. Our experiments on outdoor benchmarks demonstrate that CUBE-LLM significantly outperforms existing baselines by 21.3 points of AP-BEV on the Talk2Car dataset for 3D grounded reasoning and 17.7 points on the DriveLM dataset for complex reasoning about driving scenarios, respectively. CUBE-LLM also shows competitive results in general MLLM benchmarks such as refCOCO for 2D grounding with (87.0) average score, as well as visual question answering benchmarks such as VQAv2, GQA, SQA, POPE, etc. for complex reasoning.

1482. CausalRivers - Scaling up benchmarking of causal discovery for real-world time-series

链接：https://iclr.cc/virtual/2025/poster/27823 abstract： Causal discovery, or identifying causal relationships from observational data, is a notoriously challenging task, with numerous methods proposed to tackle it.Despite this, in-the-wild evaluation of these methods is still lacking, as works frequently rely on synthetic data evaluation and sparse real-world examples under critical theoretical assumptions.Real-world causal structures, however, are often complex, evolving over time, non-linear, and influenced by unobserved factors, makingit hard to decide on a proper causal discovery strategy.To bridge this gap, we introduce CausalRivers, the largest in-the-wild causal discovery benchmarking kit for time-series data to date.CausalRivers

features an extensive dataset on river discharge that covers the eastern German territory (666 measurement stations) and the state of Bavaria (494 measurement stations).It spans the years 2019 to 2023 with a 15-minute temporal resolution.Further, we provide additional data from a flood around the Elbe River, as an event with a pronounced distributional shift.Leveraging multiple sources of information and time-series meta-data, we constructed two distinct causal ground truth graphs (Bavaria and eastern Germany).These graphs can be sampled to generate thousands of subgraphs to benchmark causal discovery across diverse and challenging settings.To demonstrate the utility of CausalRivers, we evaluate several causal discovery approaches through a set of experiments to identify areas for improvement.CausalRivers has the potential to facilitate robust evaluations and comparisons of causal discovery methods.Besides this primary purpose, we also expect that this dataset will be relevant for connected areas of research, such as time-series forecasting and anomaly detection.Based on this, we hope to push benchmark-driven method development that fosters advanced techniques for causal discovery, as is the case for many other areas of machine learning.

1483. DenoiseVAE: Learning Molecule-Adaptive Noise Distributions for Denoising-based 3D Molecular Pre-training

链接：https://iclr.cc/virtual/2025/poster/27701 abstract：

1484. DarkBench: Benchmarking Dark Patterns in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28346 abstract： We introduce DarkBench, a comprehensive benchmark for detecting dark design patterns—manipulative techniques that influence user behavior—in interactions with large language models (LLMs). Our benchmark comprises 660 prompts across six categories: brand bias, user retention, sycophancy, anthropomorphism, harmful generation, and sneaking. We evaluate models from five leading companies (OpenAI, Anthropic, Meta, Mistral, Google) and find that some LLMs are explicitly designed to favor their developers' products and exhibit untruthful communication, among other manipulative behaviors. Companies developing LLMs should recognize and mitigate the impact of dark design patterns to promote more ethical AI.

1485. Methods with Local Steps and Random Reshuffling for Generally Smooth Non-Convex Federated Optimization

链接：https://iclr.cc/virtual/2025/poster/29511 abstract： Non-convex Machine Learning problems typically do not adhere to the standard smoothness assumption. Based on empirical findings, Zhang et al. (2020b) proposed a more realistic generalized (L_0,L_1)-smoothness assumption, though it remains largely unexplored. Many existing algorithms designed for standard smooth problems need to be revised. However, in the context of Federated Learning, only a few works address this problem but rely on additional limiting assumptions. In this paper, we address this gap in the literature: we propose and analyze new methods with local steps, partial participation of clients, and Random Reshuffling without extra restrictive assumptions beyond generalized smoothness. The proposed methods are based on the proper interplay between clients' and server's stepsizes and gradient clipping. Furthermore, we perform the first analysis of these methods under the Polyak-Łojasiewicz condition. Our theory is consistent with the known results for standard smooth problems, and our experimental results support the theoretical insights.

1486. MAST: model-agnostic sparsified training

链接：https://iclr.cc/virtual/2025/poster/28129 abstract： We introduce a novel optimization problem formulation that departs from the conventional way of minimizing machine learning model loss as a black-box function. Unlike traditional formulations, the proposed approach explicitly incorporates an initially pre-trained model and random sketch operators, allowing for sparsification of both the model and gradient during training. We establish insightful properties of the proposed objective function and highlight its connections to the standard formulation. Furthermore, we present several variants of the Stochastic Gradient Descent (SGD) method adapted to the new problem formulation, including SGD with general sampling, a distributed version, and SGD with variance reduction techniques. We achieve tighter convergence rates and relax assumptions, bridging the gap between theoretical principles and practical applications, covering several important techniques such as Dropout and Sparse training. This work presents promising opportunities to enhance the theoretical understanding of model training through a sparsification-aware optimization approach.

1487. Generating Freeform Endoskeletal Robots

链接：https://iclr.cc/virtual/2025/poster/29135 abstract： The automatic design of embodied agents (e.g. robots) has existed for 31 years and is experiencing a renaissance of interest in the literature. To date however, the field has remained narrowly focused on two kinds of anatomically simple robots: (1) fully rigid, jointed bodies; and (2) fully soft, jointless bodies. Here we bridge these two extremes with the open ended creation of terrestrial endoskeletal robots: deformable soft bodies that leverage jointed internal skeletons to move efficiently across land. Simultaneous de novo generation of external and internal structures is achieved by (i) modeling 3D endoskeletal body plans as integrated collections of elastic and rigid cells that directly attach to form soft tissues anchored to compound rigid bodies; (ii) encoding these discrete mechanical subsystems into a continuous yet coherent latent embedding; (iii) optimizing the sensorimotor coordination of each decoded design using model-free

reinforcement learning; and (iv) navigating this smooth yet highly non-convex latent manifold using evolutionary strategies. This yields an endless stream of novel species of ``higher robots'' that, like all higher animals, harness the mechanical advantages of both elastic tissues and skeletal levers for terrestrial travel. It also provides a plug-and-play experimental platform for benchmarking evolutionary design and representation learning algorithms in complex hierarchical embodied systems.

1488. Wasserstein-Regularized Conformal Prediction under General Distribution Shift

链接：https://iclr.cc/virtual/2025/poster/29180 abstract： Conformal prediction yields a prediction set with guaranteed $1-\alpha$ coverage of the true target under the i.i.d. assumption, which can fail and lead to a gap between $1-\alpha$ and the actual coverage. Prior studies bound the gap using total variation distance, which cannot identify the gap changes under distribution shift at different α, thus serving as a weak indicator of prediction set validity. Besides, existing methods are mostly limited to covariate shifts, while general joint distribution shifts are more common in practice but less researched. In response, we first propose a Wasserstein distance-based upper bound of the coverage gap and analyze the bound using probability measure pushforwards between the shifted joint data and conformal score distributions, enabling a separation of the effect of covariate and concept shifts over the coverage gap. We exploit the separation to design algorithms based on importance weighting and regularized representation learning (WR-CP) to reduce the Wasserstein bound with a finite-sample error bound. WR-CP achieves a controllable balance between conformal prediction accuracy and efficiency. Experiments on six datasets prove that WR-CP can reduce coverage gaps to 3.2% across different confidence levels and outputs prediction sets 37% smaller than the worst-case approach on average.

1489. Cut the Crap: An Economical Communication Pipeline for LLM-based Multi-Agent Systems

链接：https://iclr.cc/virtual/2025/poster/29978 abstract： Recent advancements in large language model (LLM)-powered agents have shown that collective intelligence can significantly outperform individual capabilities, largely attributed to the meticulously designed inter-agent communication topologies. Though impressive in performance, existing multi-agent pipelines inherently introduce substantial token overhead, as well as increased economic costs, which pose challenges for their large-scale deployments. In response to this challenge, we propose an economical, simple, and robust multi-agent communication framework, termed $\texttt{AgentPrune}$, which can seamlessly integrate into mainstream multi-agent systems and prunes redundant or even malicious communication messages. Technically, $\texttt{AgentPrune}$ is the first to identify and formally define the $\textit{Communication Redundancy}$ issue present in current LLM-based multi-agent pipelines, and efficiently performs one-shot pruning on the spatial-temporal message-passing graph, yielding a token-economic and high-performing communication topology.Extensive experiments across six benchmarks demonstrate that $\texttt{AgentPrune}$ $\textbf{(I)}$ achieves comparable results as state-of-the-art topologies at merely $\$5.6$ cost compared to their $\$43.7$, $\textbf{(II)}$ integrates seamlessly into existing multi-agent frameworks with $28.1\%\sim72.8\%\downarrow$ token reduction, and $\textbf{(III)}$ successfully defend against two types of agent-based adversarial attacks with $3.5\%\sim10.8\%\uparrow$ performance boost. The source code is available at \url{https://github.com/yanweiyue/AgentPrune}.

1490. You Only Prune Once: Designing Calibration-Free Model Compression With Policy Learning

链接：https://iclr.cc/virtual/2025/poster/30946 abstract： The ever-increasing size of large language models (LLMs) presents significant challenges for deployment due to their heavy computational and memory requirements. Current model pruning techniques attempt to alleviate these issues by relying heavily on external calibration datasets to determine which parameters to prune or compress, thus limiting their flexibility and scalability across different compression ratios. Moreover, these methods often cause severe performance degradation, particularly in downstream tasks, when subjected to higher compression rates. In this paper, we propose PruneNet, a novel model compression method that addresses these limitations by reformulating model pruning as a policy learning process. PruneNet decouples the pruning process from the model architecture, eliminating the need for calibration datasets. It learns a stochastic pruning policy to assess parameter importance solely based on intrinsic model properties while preserving the spectral structure to minimize information loss. PruneNet can compress the LLaMA-2-7B model in just 15 minutes, achieving over 80\% retention of its zero-shot performance with a 30\% compression ratio, outperforming existing methods that retain only 75\% performance. Furthermore, on complex multitask language understanding tasks, PruneNet demonstrates its robustness by preserving up to 80\% performance of the original model, proving itself a superior alternative to conventional structured compression techniques.

1491. PQMass: Probabilistic Assessment of the Quality of Generative Models using Probability Mass Estimation

链接：https://iclr.cc/virtual/2025/poster/28426 abstract： We propose a likelihood-free method for comparing two distributions given samples from each, with the goal of assessing the quality of generative models. The proposed approach, PQMass, provides a statistically rigorous method for assessing the performance of a single generative model or the comparison of multiple competing models. PQMass divides the sample space into non-overlapping regions and applies chi-squared tests to the number of data samples that fall within each region, giving a p-value that measures the probability that the bin counts

derived from two sets of samples are drawn from the same multinomial distribution. PQMass does not depend on assumptions regarding the density of the true distribution, nor does it rely on training or fitting any auxiliary models. We evaluate PQMass on data of various modalities and dimensions, demonstrating its effectiveness in assessing the quality, novelty, and diversity of generated samples. We further show that PQMass scales well to moderately high-dimensional data and thus obviates the need for feature extraction in practical applications.

1492. Rethinking the role of frames for SE(3)-invariant crystal structure modeling

链接：https://iclr.cc/virtual/2025/poster/28789 abstract： Crystal structure modeling with graph neural networks is essential for various applications in materials informatics, and capturing SE(3)-invariant geometric features is a fundamental requirement for these networks. A straightforward approach is to model with orientation-standardized structures through structure-aligned coordinate systems, or “frames.” However, unlike molecules, determining frames for crystal structures is challenging due to their infinite and highly symmetric nature. In particular, existing methods rely on a statically fixed frame for each structure, determined solely by its structural information, regardless of the task under consideration. Here, we rethink the role of frames, questioning whether such simplistic alignment with the structure is sufficient, and propose the concept of dynamic frames. While accommodating the infinite and symmetric nature of crystals, these frames provide each atom with a dynamic view of its local environment, focusing on actively interacting atoms. We demonstrate this concept by utilizing the attention mechanism in a recent transformer-based crystal encoder, resulting in a new architecture called CrystalFramer. Extensive experiments show that CrystalFramer outperforms conventional frames and existing crystal encoders in various crystal property prediction tasks.

1493. Iterative Label Refinement Matters More than Preference Optimization under Weak Supervision

链接：https://iclr.cc/virtual/2025/poster/28265 abstract： Language model (LM) post-training relies on two stages of human supervision: task demonstrations for supervised finetuning (SFT), followed by preference comparisons for reinforcement learning from human feedback (RLHF). As LMs become more capable, the tasks they are given become harder to supervise. Will post-training remain effective under unreliable supervision? To test this, we simulate unreliable demonstrations and comparison feedback using small LMs and time-constrained humans. We find that in the presence of unreliable supervision, SFT still retains some effectiveness, but DPO (a common RLHF algorithm) fails to improve the model beyond SFT. To address this, we propose iterative label refinement (ILR) as an alternative to RLHF. ILR improves the SFT data by using comparison feedback to decide whether human demonstrations should be replaced by model-generated alternatives, then retrains the model via SFT on the updated data. SFT+ILR outperforms SFT+DPO on several tasks with unreliable supervision (math, coding, and safe instruction-following). Our findings suggest that as LMs are used for complex tasks where human supervision is unreliable, RLHF may no longer be the best use of human comparison feedback; instead, it is better to direct feedback towards improving the training data rather than continually training the model. Our code and data are available at https://github.com/helloelwin/iterative-label-refinement.

1494. Uncovering Gaps in How Humans and LLMs Interpret Subjective Language

链接：https://iclr.cc/virtual/2025/poster/28791 abstract： Humans often rely on subjective natural language to direct language models (LLMs); for example, users might instruct the LLM to write an enthusiastic blogpost, while developers might train models to be helpful and harmless using LLM-based edits. The LLM’s operational semantics of such subjective phrases---how it adjusts its behavior when each phrase is included in the prompt---thus dictates how aligned it is with human intent. In this work, we uncover instances of misalignment between LLMs' actual operational semantics and what humans expect. Our method, TED (thesaurus error detector), first constructs a thesaurus that captures whether two phrases have similar operational semantics according to the LLM. It then elicits failures by unearthing disagreements between this thesaurus and a human-constructed reference. TED routinely produces surprising instances of misalignment; for example, Mistral 7B Instruct produces more harassing outputs when it edits text to be witty, and Llama 3 8B Instruct produces dishonest articles when instructed to make the articles enthusiastic. Our results demonstrate that humans can uncover unexpected LLM behavior by scrutinizing relationships between abstract concepts, without supervising outputs directly.

1495. Periodic Materials Generation using Text-Guided Joint Diffusion Model

链接：https://iclr.cc/virtual/2025/poster/30619 abstract： Equivariant diffusion models have emerged as the prevailing approach for generat-ing novel crystal materials due to their ability to leverage the physical symmetriesof periodic material structures. However, current models do not effectively learn thejoint distribution of atom types, fractional coordinates, and lattice structure of thecrystal material in a cohesive end-to-end diffusion framework. Also, none of thesemodels work under realistic setups, where users specify the desired characteristicsthat the generated structures must match. In this work, we introduce TGDMat, anovel text-guided diffusion model designed for 3D periodic material generation.Our approach integrates global structural knowledge through textual descriptionsat each denoising step while jointly generating atom coordinates, types, and latticestructure using a periodic-E(3)-equivariant graph neural network (GNN). Extensiveexperiments using popular datasets on benchmark tasks reveal

that TGDMat out-performs existing baseline methods by a good margin. Notably, for the structureprediction task, with just one generated sample, TGDMat outperforms all baselinemodels, highlighting the importance of text-guided diffusion. Further, in the genera-tion task, TGDMat surpasses all baselines and their text-fusion variants, showcasingthe effectiveness of the joint diffusion paradigm. Additionally, incorporating textualknowledge reduces overall training and sampling computational overhead whileenhancing generative performance when utilizing real-world textual prompts fromexperts. Code is available at https://github.com/kdmsit/TGDMat

1496. SynFlowNet: Design of Diverse and Novel Molecules with Synthesis Constraints

链接：https://iclr.cc/virtual/2025/poster/27946 abstract： Generative models see increasing use in computer-aided drug design. However, while performing well at capturing distributions of molecular motifs, they often produce synthetically inaccessible molecules. To address this, we introduce SynFlowNet, a GFlowNet model whose action space uses chemical reactions and buyable reactants to sequentially build new molecules. By incorporating forward synthesis as an explicit constraint of the generative mechanism, we aim at bridging the gap between in silico molecular generation and real world synthesis capabilities. We evaluate our approach using synthetic accessibility scores and an independent retrosynthesis tool to assess the synthesizability of our compounds, and motivate the choice of GFlowNets through considerable improvement in sample diversity compared to baselines. Additionally, we identify challenges with reaction encodings that can complicate traversal of the MDP in the backward direction. To address this, we introduce various strategies for learning the GFlowNet backward policy and thus demonstrate how additional constraints can be integrated into the GFlowNet MDP framework. This approach enables our model to successfully identify synthesis pathways for previously unseen molecules.

1497. Towards Unbiased Calibration using Meta-Regularization

链接：https://iclr.cc/virtual/2025/poster/31481 abstract： Model miscalibration has been frequently identified in modern deep neural networks. Recent work aims to improve model calibration directly through a differentiable calibration proxy. However, the calibration produced is often biased due to the binning mechanism. In this work, we propose to learn better-calibrated models via meta-regularization, which has two components: (1) gamma network (gamma-net), a meta learner that outputs sample-wise gamma value (continuous variable) for Focal loss for regularizing the backbone network; (2) smooth expected calibration error (SECE), a Gaussian-kernel based, unbiased, and differentiable surrogate to ECE that enables the smooth optimization of gamma-net. We evaluate the effectiveness of the proposed approach in regularizing neural networks towards improved and unbiased calibration on three computer vision datasets. We empirically demonstrate that: (a) learning sample-wise γ as continuous variables can effectively improve calibration; (b) SECE smoothly optimizes gamma-net towards unbiased and robust calibration with respect to the binning schemes; and (c) the combination of gamma-net and SECE achieves the best calibration performance across various calibration metrics while retaining very competitive predictive performance as compared to multiple recently proposed methods.

1498. Fast Training of Sinusoidal Neural Fields via Scaling Initialization

链接：https://iclr.cc/virtual/2025/poster/29576 abstract： Neural fields are an emerging paradigm that represent data as continuous functions parameterized by neural networks. Despite many advantages, neural fields often have a high training cost, which prevents a broader adoption. In this paper, we focus on a popular family of neural fields, called sinusoidal neural fields (SNFs), and study how it should be initialized to maximize the training speed. We find that the standard initialization scheme for SNFs---designed based on the signal propagation principle---is suboptimal. In particular, we show that by simply multiplying each weight (except for the last layer) by a constant, we can accelerate SNF training by 10$\times$. This method, coined *weight scaling*, consistently provides a significant speedup over various data domains, allowing the SNFs to train faster than more recently proposed architectures. To understand why the weight scaling works well, we conduct extensive theoretical and empirical analyses which reveal that the weight scaling not only resolves the spectral bias quite effectively but also enjoys a well-conditioned optimization trajectory.

1499. Learning the Complexity of Weakly Noisy Quantum States

链接：https://iclr.cc/virtual/2025/poster/28027 abstract： Quantifying the complexity of quantum states is a longstanding key problem in various subfields of science, ranging from quantum computing to the black-hole theory. The lower bound on quantum pure state complexity has been shown to grow linearly with system size [J. Haferkamp et al., 2022, Nat. Phys.]. However, extending this result to noisy circuit environments, which better reflect real quantum devices, remains an open challenge. In this paper, we explore the complexity of weakly noisy quantum states via the quantum learning method. We present an efficient learning algorithm, that leverages the classical shadow representation of target quantum states, to predict the circuit complexity of weakly noisy quantum states. Our algorithm is proved to be optimal in terms of sample complexity accompanied with polynomial classical processing time. Our result builds a bridge between the learning algorithm and quantum state complexity, meanwhile highlighting the power of learning algorithm in characterizing intrinsic properties of quantum states.

1500. Fast unsupervised ground metric learning with tree-Wasserstein distance

链接：https://iclr.cc/virtual/2025/poster/30359 abstract： The performance of unsupervised methods such as clustering depends on the choice of distance metric between features, or ground metric. Commonly, ground metrics are decided with heuristics or learned via supervised algorithms. However, since many interesting datasets are unlabelled, unsupervised ground metric learning approaches have been introduced. One promising option employs Wasserstein singular vectors (WSVs), which emerge when computing optimal transport distances between features and samples simultaneously. WSVs are effective, but can be prohibitively computationally expensive in some applications: $\mathcal{O}(n^2m^2(n \log(n) + m \log(m))$ for n samples and m features. In this work, we propose to augment the WSV method by embedding samples and features on trees, on which we compute the tree-Wasserstein distance (TWD). We demonstrate theoretically and empirically that the algorithm converges to a better approximation of the standard WSV approach than the best known alternatives, and does so with $\mathcal{O}(n^3+m^3+mn)$ complexity. In addition, we prove that the initial tree structure can be chosen flexibly, since tree geometry does not constrain the richness of the approximation up to the number of edge weights. This proof suggests a fast and recursive algorithm for computing the tree parameter basis set, which we find crucial to realising the efficiency gains at scale. Finally, we employ the tree-WSV algorithm to several single-cell RNA sequencing genomics datasets, demonstrating its scalability and utility for unsupervised cell-type clustering problems. These results poise unsupervised ground metric learning with TWD as a low-rank approximation of WSV with the potential for widespread application.

1501. Learning system dynamics without forgetting

链接：https://iclr.cc/virtual/2025/poster/28175 abstract： Observation-based trajectory prediction for systems with unknown dynamics is essential in fields such as physics and biology. Most existing approaches are limited to learning within a single system with fixed dynamics patterns. However, many real-world applications require learning across systems with evolving dynamics patterns, a challenge that has been largely overlooked. To address this, we systematically investigate the problem of Continual Dynamics Learning (CDL), examining task configurations and evaluating the applicability of existing techniques, while identifying key challenges. In response, we propose the Mode-switching Graph ODE (MS-GODE) model, which integrates the strengths LG-ODE and sub-network learning with a mode-switching module, enabling efficient learning over varying dynamics. Moreover, we construct a novel benchmark of biological dynamic systems for CDL, Bio-CDL, featuring diverse systems with disparate dynamics and significantly enriching the research field of machine learning for dynamic systems. Our code available at \url{https://github.com/QueuQ/MS-GODE}.

1502. QERA: an Analytical Framework for Quantization Error Reconstruction

链接：https://iclr.cc/virtual/2025/poster/30014 abstract： The growing number of parameters and computational demands of large language models (LLMs) present significant challenges for their efficient deployment.Recently, there is an increasing interest in quantizing weights to extremely low precision while offsetting the resulting error with low-rank, high-precision error reconstruction terms.The combination of quantization and low-rank approximation is now popular in both adapter-based, parameter-efficient fine-tuning methods such as LoftQ and low-precision inference techniques including ZeroQuant-V2.Usually, the low-rank terms are calculated via the singular value decomposition (SVD) of the weight quantization error,minimizing the Frobenius and spectral norms of the weight approximation error.Recent methods like LQ-LoRA and LQER introduced hand-crafted heuristics to minimize errors in layer outputs (activations) rather than weights, resulting improved quantization results.However, these heuristic methods lack an analytical solution to guide the design of quantization error reconstruction terms.In this paper, we revisit this problem and formulate an analytical framework, named Quantization Error Reconstruction Analysis (QERA),and offer a closed-form solution to the problem.We show QERA benefits both existing low-precision fine-tuning and inference methods --QERA achieves a fine-tuned accuracy gain of Δ_{acc} = 6.05\% of 2-bit RoBERTa-base on GLUE compared to LoftQ;and obtains Δ_{acc} = 2.97\% higher post-training quantization accuracy of 4-bit Llama-3.1-70B on average than ZeroQuant-V2 and Δ_{ppl} = $-$ 0.28 lower perplexity on WikiText2 than LQER.

1503. VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents

链接：https://iclr.cc/virtual/2025/poster/27679 abstract： Retrieval-augmented generation (RAG) is an effective technique that enables large language models (LLMs) to utilize external knowledge sources for generation. However, current RAG systems are solely based on text, rendering it impossible to utilize vision information like layout and images that play crucial roles in real-world multi-modality documents. In this paper, we introduce VisRAG, which tackles this issue by establishing a vision-language model (VLM)-based RAG pipeline. In this pipeline, instead of first parsing the document to obtain text, the document is directly embedded using a VLM as an image and then retrieved to enhance the generation of a VLM. Compared to traditional text-based RAG, VisRAG maximizes the retention and utilization of the data information in the original documents, eliminating the information loss introduced during the parsing process. We collect both open-source and synthetic data to train the retriever in VisRAG and explore a variety of generation methods. Experiments demonstrate that VisRAG outperforms traditional RAG in both the retrieval and generation stages, achieving a 20–40% end-to-end performance gain over traditional text-based RAG pipeline. Further analysis reveals that VisRAG is efficient in utilizing training data and demonstrates strong generalization capability, positioning it as a promising solution for RAG on multi-modality documents. Our code and data are available at https://github.com/openbmb/visrag.

1504. Self-Improvement for Neural Combinatorial Optimization: Sample Without Replacement, but Improvement

链接：https://iclr.cc/virtual/2025/poster/31479 abstract： Current methods for end-to-end constructive neural combinatorial optimization usually train a policy using behavior cloning from expert solutions or policy gradient methods from reinforcement learning. While behavior cloning is straightforward, it requires expensive expert solutions, and policy gradient methods are often computationally demanding and complex to fine-tune. In this work, we bridge the two and simplify the training process by sampling multiple solutions for random instances using the current model in each epoch and then selecting the best solution as an expert trajectory for supervised imitation learning. To achieve progressively improving solutions with minimal sampling, we introduce a method that combines round-wise Stochastic Beam Search with an update strategy derived from a provable policy improvement. This strategy refines the policy between rounds by utilizing the advantage of the sampled sequences with almost no computational overhead. We evaluate our approach on the Traveling Salesman Problem and the Capacitated Vehicle Routing Problem. The models trained with our method achieve comparable performance and generalization to those trained with expert data. Additionally, we apply our method to the Job Shop Scheduling Problem using a transformer-based architecture and outperform existing state-of-the-art methods by a wide margin.

1505. AI Sandbagging: Language Models can Strategically Underperform on Evaluations

链接：https://iclr.cc/virtual/2025/poster/30824 abstract： Trustworthy capability evaluations are crucial for ensuring the safety of AI systems, and are becoming a key component of AI regulation. However, the developers of an AI system, or the AI system itself, may have incentives for evaluations to understate the AI's actual capability. These conflicting interests lead to the problem of sandbagging – which we define as strategic underperformance on an evaluation. In this paper we assess sandbagging capabilities in contemporary language models (LMs). We prompt frontier LMs, like GPT-4 and Claude 3 Opus, to selectively underperform on dangerous capability evaluations, while maintaining performance on general (harmless) capability evaluations. Moreover, we find that models can be fine-tuned, on a synthetic dataset, to hide specific capabilities unless given a password. This behaviour generalizes to high-quality, held-out benchmarks such as WMDP. In addition, we show that both frontier and smaller models can be prompted or password-locked to target specific scores on a capability evaluation. We have mediocre success in password-locking a model to mimic the answers a weaker model would give. Overall, our results suggest that capability evaluations are vulnerable to sandbagging. This vulnerability decreases the trustworthiness of evaluations, and thereby undermines important safety decisions regarding the development and deployment of advanced AI systems.See our code at https://github.com/TeunvdWeij/sandbagging

1506. Improving Pretraining Data Using Perplexity Correlations

链接：https://iclr.cc/virtual/2025/poster/28733 abstract： Quality pretraining data is often seen as the key to high-performance language models. However, progress in understanding pretraining data has been slow due to the costly pretraining runs required for data selection experiments. We present a framework that avoids these costs and selects high-quality pretraining data without any LLM training of our own. Our work is based on a simple observation: LLM losses on many pretraining texts are correlated with downstream benchmark performance, and selecting high-correlation documents is an effective pretraining data selection method. We build a new statistical framework for data selection centered around estimates of perplexity-benchmark correlations and perform data selection using a sample of 90 LLMs taken from the Open LLM Leaderboard on texts from tens of thousands of web domains. In controlled pretraining experiments at the 160M parameter scale on 8 benchmarks, our approach outperforms DSIR on every benchmark, while matching the best data selector found in DataComp-LM, a hand-engineered bigram classifier. We have now also updated this paper to include results from preregistered experiments with new pretraining data on an aggregation of 22 benchmarks up to the 1.4B scale, showing increasing improvements of our method over others with more scale. A pip package with full documentation can be found here: https://github.com/TristanThrush/perplexity-correlations.

1507. Retri3D: 3D Neural Graphics Representation Retrieval

链接：https://iclr.cc/virtual/2025/poster/28267 abstract： Learnable 3D Neural Graphics Representations (3DNGR) have emerged as promising 3D representations for reconstructing 3D scenes from 2D images. Numerous works, including Neural Radiance Fields (NeRF), 3D Gaussian Splatting (3DGS), and their variants, have significantly enhanced the quality of these representations. The ease of construction from 2D images, suitability for online viewing/sharing, and applications in game/art design downstream tasks make it a vital 3D representation, with potential creation of large numbers of such 3D models. This necessitates large data stores, local or online, to save 3D visual data in these formats. However, no existing framework enables accurate retrieval of stored 3DNGRs. In this work, we propose, Retri3D, a framework that enables accurate and efficient retrieval of 3D scenes represented as NGRs from large data stores using text queries. We introduce a novel Neural Field Artifact Analysis technique, combined with a Smart Camera Movement Module, to select clean views and navigate pre-trained 3DNGRs. These techniques enable accurate retrieval by selecting the best viewing directions in the 3D scene for high-quality visual feature embeddings. We demonstrate that Retri3D is compatible with any NGR representation. On the LERF and ScanNet++ datasets, we show significant improvement in retrieval accuracy compared to existing techniques, while being orders of magnitude faster and storage efficient.

1508. On the Price of Differential Privacy for Hierarchical Clustering

链接：https://iclr.cc/virtual/2025/poster/27734 abstract： Hierarchical clustering is a fundamental unsupervised machine learning task with the aim of organizing data into a hierarchy of clusters. Many applications of hierarchical clustering involve sensitive user information, therefore motivating recent studies on differentially private hierarchical clustering under the rigorous

framework of Dasgupta's objective. However, it has been shown that any privacy-preserving algorithm under edge-level differential privacy necessarily suffers a large error. To capture practical applications of this problem, we focus on the weight privacy model, where each edge of the input graph is at least unit weight. We present a novel algorithm in the weight privacy model that shows significantly better approximation than known impossibility results in the edge-level DP setting. In particular, our algorithm achieves $O(\log^{1.5}n/\varepsilon)$ multiplicative error for ε-DP and runs in polynomial time, where n is the size of the input graph, and the cost is never worse than the optimal additive error in existing work. We complement our algorithm by showing if the unit-weight constraint does not apply, the lower bound for weight-level DP hierarchical clustering is essentially the same as the edge-level DP, i.e. $\Omega(n^2/\varepsilon)$ additive error. As a result, we also obtain a new lower bound of $\tilde{\Omega}(1/\varepsilon)$ additive error for balanced sparsest cuts in the weight-level DP model, which may be of independent interest. Finally, we evaluate our algorithm on synthetic and real-world datasets. Our experimental results show that our algorithm performs well in terms of extra cost and has good scalability to large graphs.

1509. CogCoM: A Visual Language Model with Chain-of-Manipulations Reasoning

链接：https://iclr.cc/virtual/2025/poster/30334 abstract： Vision-Language Models (VLMs) have shown broad effectiveness due to extensive training that aligns visual inputs with corresponding language responses. However, this conclusive alignment training causes models to overlook essential visual reasoning, leading to failures in handling detailed visual tasks and producing unfaithful responses. Drawing inspiration from human cognition in solving visual problems (e.g., marking, zoom in), this paper introduces Chain of Manipulations, a mechanism that enables VLMs to tackle problems step-by-step with evidence. After training, models can solve various visual problems by eliciting intrinsic manipulations (e.g., grounding, zoom in) with results (e.g., boxes, image) actively without relying external tools, while also allowing users to trace error causes. In this paper, we study the comprehensive methodology that includes: (1) a flexible design of manipulations based on extensive analysis, (2) an efficient automated data generation pipeline, (3) a compatible VLM architecture capable of multi-turn, multi-image, and (4) a model training process for versatile capabilities. With the design, we also manually annotate 6K high-quality samples for challenging graphical mathematical problems. Our trained model, CogCoM, equipped with this mechanism and 17B parameters, achieves SOTA performance across 9 benchmarks in 4 categories, demonstrating its effectiveness while maintaining interpretability. Code, model, and data are available at https://github.com/THUDM/CogCoM.

1510. Learning-Augmented Search Data Structures

链接：https://iclr.cc/virtual/2025/poster/29901 abstract： We study the integration of machine learning advice to improve upon traditional data structure designed for efficient search queries. Although there has been recent effort in improving the performance of binary search trees using machine learning advice, e.g., Lin et. al. (ICML 2022), the resulting constructions nevertheless suffer from inherent weaknesses of binary search trees, such as complexity of maintaining balance across multiple updates and the inability to handle partially-ordered or high-dimensional datasets. For these reasons, we focus on skip lists and KD trees in this work. Given access to a possibly erroneous oracle that outputs estimated fractional frequencies for search queries on a set of items, we construct skip lists and KD trees that provably provides the optimal expected search time, within nearly a factor of two. In fact, our learning-augmented skip lists and KD trees are still optimal up to a constant factor, even if the oracle is only accurate within a constant factor. We also demonstrate robustness by showing that our data structures achieves an expected search time that is within a constant factor of an oblivious skip list/KD tree construction even when the predictions are arbitrarily incorrect. Finally, we empirically show that our learning-augmented search data structures outperforms their corresponding traditional analogs on both synthetic and real-world datasets.

1511. MixMax: Distributional Robustness in Function Space via Optimal Data Mixtures

链接：https://iclr.cc/virtual/2025/poster/28992 abstract： Machine learning models are often required to perform well across several pre-defined settings, such as a set of user groups. Worst-case performance is a common metric to capture this requirement, and is the objective of group distributionally robust optimization (group DRO). Unfortunately, these methods struggle when the loss is non-convex in the parameters, or the model class is non-parametric. Here, we make a classical move to address this: we reparameterize group DRO from parameter space to function space, which results in a number of advantages. First, we show that group DRO over the space of bounded functions admits a minimax theorem. Second, for cross-entropy and mean squared error, we show that the minimax optimal mixture distribution is the solution of a simple convex optimization problem. Thus, provided one is working with a model class of universal function approximators, group DRO can be solved by a convex optimization problem followed by a classical risk minimization problem. We call our method MixMax. In our experiments, we found that MixMax matched or outperformed the standard group DRO baselines, and in particular, MixMax improved the performance of XGBoost over the only baseline, data balancing, for variations of the ACSIncome and CelebA annotations datasets.

1512. AdaWM: Adaptive World Model based Planning for Autonomous Driving

链接：https://iclr.cc/virtual/2025/poster/29891 abstract： World model based reinforcement learning (RL) has emerged as a

promising approach for autonomous driving, which learns a latent dynamics model and uses it to train a planning policy. To speed up the learning process, the pretrain-finetune paradigm is often used, where online RL is initialized by a pretrained model and a policy learned offline. However, naively performing such initialization in RL may result in dramatic performance degradation during the online interactions in the new task. To tackle this challenge, we first analyze the performance degradation and identify two primary root causes therein: the mismatch of the planning policy and the mismatch of the dynamics model, due to distribution shift. We further analyze the effects of these factors on performance degradation during finetuning, and our findings reveal that the choice of finetuning strategies plays a pivotal role in mitigating these effects. We then introduce AdaWM, an Adaptive World Model based planning method, featuring two key steps: (a) mismatch identification, which quantifies the mismatches and informs the finetuning strategy, and (b) alignment-driven finetuning, which selectively updates either the policy or the model as needed using efficient low-rank updates. Extensive experiments on the challenging CARLA driving tasks demonstrate that AdaWM significantly improves the finetuning process, resulting in more robust and efficient performance in autonomous driving systems.

1513. HASARD: A Benchmark for Vision-Based Safe Reinforcement Learning in Embodied Agents

链接：https://iclr.cc/virtual/2025/poster/30963 abstract：

1514. Toward Understanding In-context vs. In-weight Learning

链接：https://iclr.cc/virtual/2025/poster/29178 abstract：

1515. LoRA Done RITE: Robust Invariant Transformation Equilibration for LoRA Optimization

链接：https://iclr.cc/virtual/2025/poster/29397 abstract： Low-rank adaption (LoRA) is a widely used parameter-efficient finetuning method for LLM that reduces memory requirements. However, current LoRA optimizers lack transformation invariance, meaning the updates depending on how the two LoRA factors are scaled or rotated. This deficiency leads to inefficient learning and sub-optimal solutions in practice. This paper introduces LoRA-RITE, a novel adaptive matrix preconditioning method for LoRA optimization, which can achieve transformation invariance and remain computationally efficient. We provide theoretical analysis to demonstrate the benefit of our method and conduct experiments on various LLM tasks with different models including Gemma 2B, 7B, and mT5-XXL. The results demonstrate consistent improvements against existing optimizers. For example, replacing Adam with LoRA-RITE during LoRA fine-tuning of Gemma-2B yielded 4.6% accuracy gain on Super-Natural Instructions and 3.5% accuracy gain across other four LLM benchmarks (HellaSwag, ArcChallenge, GSM8K, OpenBookQA).

1516. Improving Unsupervised Constituency Parsing via Maximizing Semantic Information

链接：https://iclr.cc/virtual/2025/poster/28229 abstract： Unsupervised constituency parsers organize phrases within a sentence into a tree-shaped syntactic constituent structure that reflects the organization of sentence semantics. However, the traditional objective of maximizing sentence log-likelihood (LL) does not explicitly account for the close relationship between the constituent structure and the semantics, resulting in a weak correlation between LL values and parsing accuracy.In this paper, we introduce a novel objective that trains parsers by maximizing SemInfo, the semantic information encoded in constituent structures.We introduce a bag-of-substrings model to represent the semantics and estimate the SemInfo value using the probability-weighted information metric.We apply the SemInfo maximization objective to training Probabilistic Context-Free Grammar (PCFG) parsers and develop a Tree Conditional Random Field (TreeCRF)-based model to facilitate the training. Experiments show that SemInfo correlates more strongly with parsing accuracy than LL, establishing SemInfo as a better unsupervised parsing objective.As a result, our algorithm significantly improves parsing accuracy by an average of 7.85 sentence-F1 scores across five PCFG variants and in four languages, achieving state-of-the-art level results in three of the four languages.

1517. GIFT: Unlocking Full Potential of Labels in Distilled Dataset at Near-zero Cost

链接：https://iclr.cc/virtual/2025/poster/30327 abstract： Recent advancements in dataset distillation have demonstrated the significant benefits of employing soft labels generated by pre-trained teacher models. In this paper, we introduce a novel perspective by emphasizing the full utilization of labels. We first conduct a comprehensive comparison of various loss functions for soft label utilization in dataset distillation, revealing that the model trained on the synthetic dataset exhibits high sensitivity to the choice of loss function for soft label utilization. This finding highlights the necessity of a universal loss function for training models on synthetic datasets. Building on these insights, we introduce an extremely simple yet surprisingly effective plug-and-play approach, GIFT, which encompasses soft label refinement and a cosine similarity-based loss function to efficiently leverage full label information. Extensive experiments indicate that GIFT consistently enhances state-of-the-art dataset distillation methods

across various dataset scales without incurring additional computational costs. Importantly, GIFT significantly enhances cross-optimizer generalization, an area previously overlooked. For instance, on ImageNet-1K with IPC = 10, GIFT enhances the state-of-the-art method RDED by 30.8% in cross-optimizer generalization. Our code is available at https://github.com/LINs-lab/GIFT.

1518. CoMRes: Semi-Supervised Time Series Forecasting Utilizing Consensus Promotion of Multi-Resolution

链接：https://iclr.cc/virtual/2025/poster/29110 abstract： Long-term time series forecasting poses significant challenges due to the complex dynamics and temporal variations, particularly when dealing with unseen patterns and data scarcity. Traditional supervised learning approaches, which rely on cleaned and labeled data, struggle to capture these unseen characteristics, limiting their effectiveness in real-world applications. In this study, we propose a semi-supervised approach that leverages multi-view setting on augmented data without requiring explicit future values as labels to address these limitations. By introducing a consensus promotion framework, our method enhances agreement among multiple single-view models on unseen augmented data. This approach not only improves forecasting accuracy but also mitigates error accumulation in long-horizon predictions. Furthermore, we explore the impact of autoregressive and non-autoregressive decoding schemes on error propagation, demonstrating the robustness of our model in extending prediction horizons. Experimental results show that our proposed method not only surpasses traditional supervised models in accuracy but also exhibits greater robustness when extending the prediction horizon.

1519. Drop-Upcycling: Training Sparse Mixture of Experts with Partial Re-initialization

链接：https://iclr.cc/virtual/2025/poster/28794 abstract： The Mixture of Experts (MoE) architecture reduces the training and inference cost significantly compared to a dense model of equivalent capacity. Upcycling is an approach that initializes and trains an MoE model using a pre-trained dense model. While upcycling leads to initial performance gains, the training progresses slower than when trained from scratch, leading to suboptimal performance in the long term. We propose Drop-Upcycling - a method that effectively addresses this problem. Drop-Upcycling combines two seemingly contradictory approaches: utilizing the knowledge of pre-trained dense models while statistically re-initializing some parts of the weights. This approach strategically promotes expert specialization, significantly enhancing the MoE model's efficiency in knowledge acquisition. Extensive large-scale experiments demonstrate that Drop-Upcycling significantly outperforms previous MoE construction methods in the long term, specifically when training on hundreds of billions of tokens or more.As a result, our MoE model with 5.9B active parameters achieves comparable performance to a 13B dense model in the same model family, while requiring approximately 1/4 of the training FLOPs.All experimental resources, including source code, training data, model checkpoints and logs, are publicly available to promote reproducibility and future research on MoE.

1520. Optimal Transport for Time Series Imputation

链接：https://iclr.cc/virtual/2025/poster/27784 abstract： Missing data imputation through distribution alignment has demonstrated advantages for non-temporal datasets but exhibits suboptimal performance in time-series applications. The primary obstacle is crafting a discrepancy measure that simultaneously (1) captures temporal patterns—accounting for periodicity and temporal dependencies inherent in time-series—and (2) accommodates non-stationarity, ensuring robustness amidst multiple coexisting temporal patterns. In response to these challenges, we introduce the Proximal Spectrum Wasserstein (PSW) discrepancy, a novel discrepancy tailored for comparing two \textit{sets} of time-series based on optimal transport. It incorporates a pairwise spectral distance to encapsulate temporal patterns, and a selective matching regularization to accommodate non-stationarity. Subsequently, we develop the PSW for Imputation (PSW-I) framework, which iteratively refines imputation results by minimizing the PSW discrepancy. Extensive experiments demonstrate that PSW-I effectively accommodates temporal patterns and non-stationarity, outperforming prevailing time-series imputation methods. Code is available at https://github.com/FMLYD/PSW-I.

1521. Is Large-scale Pretraining the Secret to Good Domain Generalization?

链接：https://iclr.cc/virtual/2025/poster/27866 abstract： Multi-Source Domain Generalization (DG) is the task of training on multiple source domains and achieving high classification performance on unseen target domains. Recent methods combine robust features from web-scale pretrained backbones with new features learned from source data, and this has dramatically improved benchmark results. However, it remains unclear if DG finetuning methods are becoming better over time, or if improved benchmark performance is simply an artifact of stronger pre-training. Prior studies have shown that perceptual similarity to pre-training data correlates with zero-shot performance, but we find the effect limited in the DG setting. Instead, we posit that having perceptually similar data in pretraining is not enough; and that it is how well these data were learned that determines performance. This leads us to introduce the Alignment Hypothesis, which states that the final DG performance will be high if and only if alignment of image and class label text embeddings is high. Our experiments confirm the Alignment Hypothesis is true, and we use it as an analysis tool of existing DG methods evaluated on DomainBed datasets by splitting evaluation data into In-pretraining (IP) and Out-of-pretraining (OOP). We show that all evaluated DG methods struggle on DomainBed-OOP, while recent methods excel on DomainBed-IP. Put together, our findings highlight the need for DG methods which can generalize beyond pretraining alignment. We release DomainBed-OOP at https://huggingface.co/datasets/PTeterwak/DomainBed_OOP.

1522. Cross-Domain Off-Policy Evaluation and Learning for Contextual Bandits

链接：https://iclr.cc/virtual/2025/poster/29233 abstract： Off-Policy Evaluation and Learning (OPE/L) in contextual bandits is rapidly gaining popularity in real systems because new policies can be evaluated and learned securely using only historical logged data. However, existing methods in OPE/L cannot handle many challenging but prevalent scenarios such as few-shot data, deterministic logging policies, and new actions. In many applications, such as personalized medicine, content recommendations, education, and advertising, we need to evaluate and learn new policies in the presence of these challenges. Existing methods cannot evaluate and optimize effectively in these situations due to the notorious variance issue or limited exploration in the logged data. To enable OPE/L even under these unsolved challenges, we propose a new problem setup of Cross-Domain OPE/L, where we have access not only to the logged data from the target domain in which the new policy will be implemented but also to logged datasets collected from other domains. This novel formulation is widely applicable because we can often use historical data not only from the target hospital, country, device, or user segment but also from other hospitals, countries, devices, or segments. We develop a new estimator and policy gradient method to solve OPE/L by leveraging both target and source datasets, resulting in substantially enhanced OPE/L in the previously unsolved situations in our empirical evaluations.

1523. Multimodal Unsupervised Domain Generalization by Retrieving Across the Modality Gap

链接：https://iclr.cc/virtual/2025/poster/29088 abstract： Domain generalization (DG) is an important problem that learns a model which generalizes to unseen test domains leveraging one or more source domains, under the assumption of shared label spaces. However, most DG methods assume access to abundant source data in the target label space, a requirement that proves overly stringent for numerous real-world applications, where acquiring the same label space as the target task is prohibitively expensive. For this setting, we tackle the multimodal version of the unsupervised domain generalization (MUDG) problem, which uses a large task-agnostic unlabeled source dataset during finetuning. Our framework does not explicitly assume any relationship between the source dataset and target task. Instead, it relies only on the premise that the source dataset can be accurately and efficiently searched in a joint vision-language space. We make three contributions in the MUDG setting. Firstly, we show theoretically that cross-modal approximate nearest neighbor search suffers from low recall due to the large distance between text queries and the image centroids used for coarse quantization. Accordingly, we propose paired k-means, a simple clustering algorithm that improves nearest neighbor recall by storing centroids in query space instead of image space. Secondly, we propose an adaptive text augmentation scheme for target labels designed to improve zero-shot accuracy and diversify retrieved image data. Lastly, we present two simple but effective components to further improve downstream target accuracy. We compare against state-of-the-art name-only transfer, source-free DG and zero-shot (ZS) methods on their respective benchmarks and show consistent improvement in accuracy on 20 diverse datasets. Code is available: https://github.com/Chris210634/mudg

1524. ANaGRAM: A Natural Gradient Relative to Adapted Model for efficient PINNs learning

链接：https://iclr.cc/virtual/2025/poster/28381 abstract：

1525. Systematic Outliers in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28203 abstract： Outliers have been widely observed in Large Language Models (LLMs), significantly impacting model performance and posing challenges for model compression. Understanding the functionality and formation mechanisms of these outliers is critically important. Existing works, however, largely focus on reducing the impact of outliers from an algorithmic perspective, lacking an in-depth investigation into their causes and roles. In this work, we provide a detailed analysis of the formation process, underlying causes, and functions of outliers in LLMs. We define and categorize three types of outliers—activation outliers, weight outliers, and attention outliers—and analyze their distributions across different dimensions, uncovering inherent connections between their occurrences and their ultimate influence on the attention mechanism. Based on these observations, we hypothesize and explore the mechanisms by which these outliers arise and function, demonstrating through theoretical derivations and experiments that they emerge due to the self-attention mechanism's softmax operation. These outliers act as implicit context-aware scaling factors within the attention mechanism. As these outliers stem from systematic influences, we term them systematic outliers. Our study not only enhances the understanding of Transformer-based LLMs but also shows that structurally eliminating outliers can accelerate convergence and improve model compression. The code is avilable at \url{https://github.com/an-yongqi/systematic-outliers}.

1526. Fantastic Targets for Concept Erasure in Diffusion Models and Where To Find Them

链接：https://iclr.cc/virtual/2025/poster/28045 abstract：

1527. Needle Threading: Can LLMs Follow Threads Through Near-Million-Scale Haystacks?

链接：https://iclr.cc/virtual/2025/poster/27857 abstract： As the context limits of Large Language Models (LLMs) increase, the range ofpossible applications and downstream functions broadens. In many real-worldtasks, decisions depend on details scattered across collections of often disparatedocuments containing mostly irrelevant information. Long-context LLMs appearwell-suited to this form of complex information retrieval and reasoning, which hastraditionally proven costly and time-consuming. However, although the development of longer context models has seen rapid gains in recent years, our understanding of how effectively LLMs use their context has not kept pace. To addressthis, we conduct a set of retrieval experiments designed to evaluate the capabilitiesof 17 leading LLMs, such as their ability to follow threads of information throughthe context window. Strikingly, we find that many models are remarkably thread-safe: capable of simultaneously following multiple threads without significant lossin performance. Still, for many models, we find the effective context limit is significantly shorter than the supported context length, with accuracy decreasing asthe context window grows. Our study also highlights the important point that token counts from different tokenizers should not be directly compared—they oftencorrespond to substantially different numbers of written characters. We releaseour code and long context experimental data.

1528. Residual Connections and Normalization Can Provably Prevent Oversmoothing in GNNs

链接：https://iclr.cc/virtual/2025/poster/28713 abstract： Residual connections and normalization layers have become standard design choices for graph neural networks (GNNs), and were proposed as solutions to the mitigate the oversmoothing problem in GNNs. However, how exactly these methods help alleviate the oversmoothing problem from a theoretical perspective is not well understood. In this work, we provide a formal and precise characterization of (linearized) GNNs with residual connections and normalization layers. We establish that (a) for residual connections, the incorporation of the initial features at each layer can prevent the signal from becoming too smooth, and determines the subspace of possible node representations; (b) batch normalization prevents a complete collapse of the output embedding space to a one-dimensional subspace through the individual rescaling of each column of the feature matrix. This results in the convergence of node representations to the top-k eigenspace of the message-passing operator; (c) moreover, we show that the centering step of a normalization layer — which can be understood as a projection — alters the graph signal in message-passing in such a way that relevant information can become harder to extract. Building on the last theoretical insight, we introduce GraphNormv2, a novel and principled normalization layer. GraphNormv2 features a learnable centering step designed to preserve the integrity of the original graph signal. Experimental results corroborate the effectiveness of our method, demonstrating improved performance across various GNN architectures and tasks.

1529. Gaussian Mixture Counterfactual Generator

链接：https://iclr.cc/virtual/2025/poster/28533 abstract： We address the individualized treatment effect (ITE) estimation problem, focusing on continuous, multidimensional, and time-dependent treatments for precision medicine. The central challenge lies in modeling these complex treatment scenarios while capturing dynamic patient responses and minimizing reliance on control data. We propose the Gaussian Mixture Counterfactual Generator (GMCG), a generative model that transforms the Gaussian mixture model—traditionally a tool for clustering and density estimation—into a new tool explicitly geared toward causal inference. This approach generates robust counterfactuals by effectively handling continuous and multidimensional treatment spaces. We evaluate GMCG on synthetic crossover trial data and simulated datasets, demonstrating its superior performance over existing methods, particularly in scenarios with limited control data. GMCG derives its effectiveness from modeling the joint distribution of covariates, treatments, and outcomes using a latent state vector while employing a conditional distribution of the state vector to suppress confounding and isolate treatment-outcome relationships.

1530. Efficient Top-m Data Values Identification for Data Selection

链接：https://iclr.cc/virtual/2025/poster/28522 abstract： Data valuation has found many real-world applications, e.g., data pricing and data selection. However, the most adopted approach -- Shapley value (SV) -- is computationally expensive due to the large number of model trainings required. Fortunately, most applications (e.g., data selection) require only knowing the m data points with the highest data values (i.e., top-m data values), which implies the potential for fewer model trainings as exact data values are not required. Existing work formulates top-m Shapley value identification as top-m arms identification in multi-armed bandits (MAB). However, the proposed approach falls short because it does not utilize data features to predict data values, a method that has been shown empirically to be effective. A recent top-m arms identification work does consider the use of arm features while assuming a linear relationship between arm features and rewards, which is often not satisfied in data valuation. To this end, we propose the GPGapE algorithm that uses the Gaussian process to model the \emph{non-linear} mapping from data features to data values, removing the linear assumption. We theoretically analyze the correctness and stopping iteration of GPGapE in finding an (ϵ, δ)-approximation to the top-m data values. We further improve the computational efficiency, by calculating data values using small data subsets to reduce the computation cost of model training. We empirically demonstrate that GPGapE outperforms other baselines in top-m data values identification, noisy data detection, and data subset selection on real-world datasets. We also demonstrate the efficiency of our GPGapE in data selection for large language model fine-tuning.

1531. Gradient correlation is a key ingredient to accelerate SGD with momentum

链接：https://iclr.cc/virtual/2025/poster/31139 abstract： Empirically, it has been observed that adding momentum to Stochastic Gradient Descent (SGD) accelerates the convergence of the algorithm.However, the literature has been rather pessimistic, even in the case of convex functions, about the possibility of theoretically proving this observation.We investigate the possibility of obtaining accelerated convergence of the Stochastic Nesterov Accelerated Gradient (SNAG), a momentum-based version of SGD, when minimizing a sum of functions in a convex setting. We demonstrate that the average correlation between gradients allows to verify the strong growth condition, which is the key ingredient to obtain acceleration with SNAG.Numerical experiments, both in linear regression and deep neural network optimization, confirm in practice our theoretical results.

1532. Learning to Search from Demonstration Sequences

链接：https://iclr.cc/virtual/2025/poster/27934 abstract： Search and planning are essential for solving many real-world problems. However, in numerous learning scenarios, only action-observation sequences, such as demonstrations or instruction sequences, are available for learning. Relying solely on supervised learning with these sequences can lead to sub-optimal performance due to the vast, unseen search space encountered during training. In this paper, we introduce Differentiable Tree Search Network (D-TSN), a novel neural network architecture that learns to construct search trees from just sequences of demonstrations by performing gradient descent on a best-first search tree construction algorithm. D-TSN enables the joint learning of submodules, including an encoder, value function, and world model, which are essential for planning. To construct the search tree, we employ a stochastic tree expansion policy and formulate it as another decision-making task. Then, we optimize the tree expansion policy via REINFORCE with an effective variance reduction technique for the gradient computation. D-TSN can be applied to problems with a known world model or to scenarios where it needs to jointly learn a world model with a latent state space. We study problems from these two scenarios, including Game of 24, 2D grid navigation, and Procgen games, to understand when D-TSN is more helpful. Through our experiments, we show that D-TSN is effective, especially when the world model with a latent state space is jointly learned. The code is available at https://github.com/dixantmittal/differentiable-tree-search-network.

1533. An Information Criterion for Controlled Disentanglement of Multimodal Data

链接：https://iclr.cc/virtual/2025/poster/31055 abstract： Multimodal representation learning seeks to relate and decompose information inherent in multiple modalities. By disentangling modality-specific information from information that is shared across modalities, we can improve interpretability and robustness and enable downstream tasks such as the generation of counterfactual outcomes. Separating the two types of information is challenging since they are often deeply entangled in many real-world applications. We propose $\textbf{Disentangled}$ $\textbf{S}$elf-$\textbf{S}$upervised $\textbf{L}$earning (DisentangledSSL), a novel self-supervised approach for learning disentangled representations. We present a comprehensive analysis of the optimality of each disentangled representation, particularly focusing on the scenario not covered in prior work where the so-called $\textit{Minimum Necessary Information}$ (MNI) point is not attainable. We demonstrate that \algo successfully learns shared and modality-specific features on multiple synthetic and real-world datasets and consistently outperforms baselines on various downstream tasks, including prediction tasks for vision-language data, as well as molecule-phenotype retrieval tasks for biological data.

1534. Mitigating Parameter Interference in Model Merging via Sharpness-Aware Fine-Tuning

链接：https://iclr.cc/virtual/2025/poster/28919 abstract： Large-scale deep learning models with a pretraining-finetuning paradigm have led to a surge of numerous task-specific models fine-tuned from a common pre-trained model.Recently, several research efforts have been made on merging these large models into a single multi-task model, particularly with simple arithmetic on parameters.Such merging methodology faces a central challenge: interference between model parameters fine-tuned on different tasks.Few recent works have focused on designing a new fine-tuning scheme that can lead to small parameter interference, however at the cost of the performance of each task-specific fine-tuned model and thereby limiting that of a merged model.To improve the performance of a merged model, we note that a fine-tuning scheme should aim for (1) smaller parameter interference and (2) better performance of each fine-tuned model on the corresponding task.In this work, we aim to design a new fine-tuning objective function to work towards these two goals.In the course of this process, we find such objective function to be strikingly similar to sharpness-aware minimization (SAM) objective function, which aims to achieve generalization by finding flat minima.Drawing upon our observation, we propose to fine-tune pre-trained models via sharpness-aware minimization.The experimental and theoretical results showcase the effectiveness and orthogonality of our proposed approach, improving performance upon various merging and fine-tuning methods.Our code is available at https://github.com/baiklab/SAFT-Merge.

1535. Language Models Are Implicitly Continuous

链接：https://iclr.cc/virtual/2025/poster/29613 abstract： Language is typically modelled with discrete sequences. However, the most successful approaches to language modelling, namely neural networks, are continuous and smooth function

approximators.In this work, we show that Transformer-based language models implicitly learn to represent sentences as continuous-time functions defined over a continuous input space. This phenomenon occurs in most state-of-the-art Large Language Models (LLMs), including Llama2, Llama3, Phi3, Gemma, Gemma2, and Mistral, and suggests that LLMs reason about language in ways that fundamentally differ from humans.Our work formally extends Transformers to capture the nuances of time and space continuity in both input and output space.Our results challenge the traditional interpretation of how LLMs understand language, with several linguistic and engineering implications.

1536. Weighted Point Set Embedding for Multimodal Contrastive Learning Toward Optimal Similarity Metric

链接：https://iclr.cc/virtual/2025/poster/27976 abstract： In typical multimodal contrastive learning, such as CLIP, encoders produce onepoint in the latent representation space for each input. However, one-point representationhas difficulty in capturing the relationship and the similarity structure of ahuge amount of instances in the real world. For richer classes of the similarity, wepropose the use of weighted point sets, namely, sets of pairs of weight and vector,as representations of instances. In this work, we theoretically show the benefitof our proposed method through a new understanding of the contrastive loss ofCLIP, which we call symmetric InfoNCE. We clarify that the optimal similaritythat minimizes symmetric InfoNCE is the pointwise mutual information, and showan upper bound of excess risk on downstream classification tasks of representationsthat achieve the optimal similarity. In addition, we show that our proposedsimilarity based on weighted point sets consistently achieves the optimal similarity.To verify the effectiveness of our proposed method, we demonstrate pretraining oftext-image representation models and classification tasks on common benchmarks.

1537. PABBO: Preferential Amortized Black-Box Optimization

链接：https://iclr.cc/virtual/2025/poster/29251 abstract： Preferential Bayesian Optimization (PBO) is a sample-efficient method to learn latent user utilities from preferential feedback over a pair of designs. It relies on a statistical surrogate model for the latent function, usually a Gaussian process, and an acquisition strategy to select the next candidate pair to get user feedback on. Due to the non-conjugacy of the associated likelihood, every PBO step requires a significant amount of computations with various approximate inference techniques. This computational overhead is incompatible with the way humans interact with computers, hindering the use of PBO in real-world cases. Building on the recent advances of amortized BO, we propose to circumvent this issue by fully amortizing PBO, meta-learning both the surrogate and the acquisition function. Our method comprises a novel transformer neural process architecture, trained using reinforcement learning and tailored auxiliary losses.On a benchmark composed of synthetic and real-world datasets, our method is several orders of magnitude faster than the usual Gaussian process-based strategies and often outperforms them in accuracy.

1538. Tighter Privacy Auditing of DP-SGD in the Hidden State Threat Model

链接：https://iclr.cc/virtual/2025/poster/27756 abstract： Machine learning models can be trained with formal privacy guarantees via differentially private optimizers such as DP-SGD. In this work, we focus on a threat model where the adversary has access only to the final model, with no visibility into intermediate updates. In the literature, this ``hidden state'' threat model exhibits a significant gap between the lower bound from empirical privacy auditing and the theoretical upper bound provided by privacy accounting. To challenge this gap, we propose to audit this threat model with adversaries that craft a gradient sequence designed to maximize the privacy loss of the final model without relying on intermediate updates. Our experiments show that this approach consistently outperforms previous attempts at auditing the hidden state model. Furthermore, our results advance the understanding of achievable privacy guarantees within this threat model. Specifically, when the crafted gradient is inserted at every optimization step, we show that concealing the intermediate model updates in DP-SGD does not enhance the privacy guarantees. The situation is more complex when the crafted gradient is not inserted at every step: our auditing lower bound matches the privacy upper bound only for an adversarially-chosen loss landscape and a sufficiently large batch size. This suggests that existing privacy upper bounds can be improved in certain regimes.

1539. Jump Your Steps: Optimizing Sampling Schedule of Discrete Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/28310 abstract： Diffusion models have seen notable success in continuous domains, leading to the development of discrete diffusion models (DDMs) for discrete variables. Despite recent advances, DDMs face the challenge of slow sampling speeds. While parallel sampling methods like τ-leaping accelerate this process, they introduce *Compounding Decoding Error* (CDE), where discrepancies arise between the true distribution and the approximation from parallel token generation, leading to degraded sample quality. In this work, we present *Jump Your Steps* (JYS), a novel approach that optimizes the allocation of discrete sampling timesteps by minimizing CDE without extra computational cost. More precisely, we derive a practical upper bound on CDE and propose an efficient algorithm for searching for the optimal sampling schedule. Extensive experiments across image, music, and text generation show that JYS significantly improves sampling quality, establishing it as a versatile framework for enhancing DDM performance for fast sampling.

1540. PRDP: Progressively Refined Differentiable Physics

链接：https://iclr.cc/virtual/2025/poster/30710 abstract： The physics solvers employed for neural network training are primarily iterative, and hence, differentiating through them introduces a severe computational burden as iterations grow large. Inspired by works in bilevel optimization, we show that full accuracy of the network is achievable through physics significantly coarser than fully converged solvers. We propose progressively refined differentiable physics (PRDP), an approach that identifies the level of physics refinement sufficient for full training accuracy. By beginning with coarse physics, adaptively refining it during training, and stopping refinement at the level adequate for training, it enables significant compute savings without sacrificing network accuracy. Our focus is on differentiating iterative linear solvers for sparsely discretized differential operators, which are fundamental to scientific computing. PRDP is applicable to both unrolled and implicit differentiation. We validate its performance on a variety of learning scenarios involving differentiable physics solvers such as inverse problems, autoregressive neural emulators, and correction-based neural-hybrid solvers. In the challenging example of emulating the Navier-Stokes equations, we reduce training time by 62%.

1541. TweedieMix: Improving Multi-Concept Fusion for Diffusion-based Image/Video Generation

链接：https://iclr.cc/virtual/2025/poster/28914 abstract： Despite significant advancements in customizing text-to-image and video generation models, generating images and videos that effectively integrate multiple personalized concepts remains challenging. To address this, we present TweedieMix, a novel method for composing customized diffusion models during the inference phase. By analyzing the properties of reverse diffusion sampling, our approach divides the sampling process into two stages. During the initial steps, we apply a multiple object-aware sampling technique to ensure the inclusion of the desired target objects. In the later steps, we blend the appearances of the custom concepts in the de-noised image space using Tweedie's formula. Our results demonstrate that TweedieMix can generate multiple personalized concepts with higher fidelity than existing methods. Moreover, our framework can be effortlessly extended to image-to-video diffusion models by extending the residual layer's features across frames}, enabling the generation of videos that feature multiple personalized concepts.

1542. Generalized Consistency Trajectory Models for Image Manipulation

链接：https://iclr.cc/virtual/2025/poster/29202 abstract： Diffusion-based generative models excel in unconditional generation, as well as on applied tasks such as image editing and restoration. The success of diffusion models lies in the iterative nature of diffusion: diffusion breaks down the complex process of mapping noise to data into a sequence of simple denoising tasks. Moreover, we are able to exert fine-grained control over the generation process by injecting guidance terms into each denoising step. However, the iterative process is also computationally intensive, often taking from tens up to thousands of function evaluations. Although consistency trajectory models (CTMs) enable traversal between any time points along the probability flow ODE (PFODE) and score inference with a single function evaluation, CTMs only allow translation from Gaussian noise to data. Thus, this work aims to unlock the full potential of CTMs by proposing generalized CTMs (GCTMs), which translate between arbitrary distributions via ODEs. We discuss the design space of GCTMs and demonstrate their efficacy in various image manipulation tasks such as image-to-image translation, restoration, and editing. Code is available at https://github.com/1202kbs/GCTM.

1543. Dream to Manipulate: Compositional World Models Empowering Robot Imitation Learning with Imagination

链接：https://iclr.cc/virtual/2025/poster/31075 abstract： A world model provides an agent with a representation of its environment, enabling it to predict the causal consequences of its actions. Current world models typically cannot directly and explicitly imitate the actual environment in front of a robot, often resulting in unrealistic behaviors and hallucinations that make them unsuitable for real-world robotics applications. To overcome those challenges, we propose to rethink robot world models as learnable digital twins. We introduce DreMa, a new approach for constructing digital twins automatically using learned explicit representations of the real world and its dynamics, bridging the gap between traditional digital twins and world models.DreMa replicates the observed world and its structure by integrating Gaussian Splatting and physics simulators, allowing robots to imagine novel configurations of objects and to predict the future consequences of robot actions thanks to its compositionality.We leverage this capability to generate new data for imitation learning by applying equivariant transformations to a small set of demonstrations. Our evaluations across various settings demonstrate significant improvements in accuracy and robustness by incrementing actions and object distributions, reducing the data needed to learn a policy and improving the generalization of the agents. As a highlight, we show that a real Franka Emika Panda robot, powered by DreMa’s imagination, can successfully learn novel physical tasks from just a single example per task variation (one-shot policy learning).Our project page can be found in: https://dreamtomanipulate.github.io/.

1544. NeurFlow: Interpreting Neural Networks through Neuron Groups and Functional Interactions

链接：https://iclr.cc/virtual/2025/poster/30270 abstract： Understanding the inner workings of neural networks is essential for enhancing model performance and interpretability. Current research predominantly focuses on examining the connection between individual neurons and the model's final predictions, which suffers from challenges in interpreting the internal workings of the model, particularly when neurons encode multiple unrelated features. In this paper, we propose a novel framework that

transitions the focus from analyzing individual neurons to investigating groups of neurons, shifting the emphasis from neuron-output relationships to the functional interactions between neurons. Our automated framework, NeurFlow, first identifies core neurons and clusters them into groups based on shared functional relationships, enabling a more coherent and interpretable view of the network's internal processes. This approach facilitates the construction of a hierarchical circuit representing neuron interactions across layers, thus improving interpretability while reducing computational costs. Our extensive empirical studies validate the fidelity of our proposed NeurFlow. Additionally, we showcase its utility in practical applications such as image debugging and automatic concept labeling, thereby highlighting its potential to advance the field of neural network explainability.

1545. HyperPLR: Hypergraph Generation through Projection, Learning, and Reconstruction

链接：https://iclr.cc/virtual/2025/poster/29533 abstract： Hypergraphs are essential in modeling higher-order complex networks, excelling in representing group interactions within real-world contexts. This is particularly evident in collaboration networks, where they facilitate the capture of groupwise polyadic patterns, extending beyond traditional pairwise dyadic interactions. The use of hypergraph generators, or generative models, is a crucial method for promoting and validating our understanding of these structures. If such generators accurately replicate observed hypergraph patterns, it reinforces the validity of our interpretations. In this context, we introduce a novel hypergraph generative paradigm, HyperPLR, encompassing three phases: Projection, Learning, and Reconstruction. Initially, the hypergraph is projected onto a weighted graph. Subsequently, the model learns this graph's structure within a latent space, while simultaneously computing a distribution between the hyperedge and the projected graph. Finally, leveraging the learned model and distribution, HyperPLR generates new weighted graphs and samples cliques from them. These cliques are then used to reconstruct new hypergraphs by solving a specific clique cover problem.We have evaluated HyperPLR on existing real-world hypergraph datasets, which consistently demonstrate superior performance and validate the effectiveness of our approach.

1546. Semi-Parametric Retrieval via Binary Bag-of-Tokens Index

链接：https://iclr.cc/virtual/2025/poster/28545 abstract： Information retrieval has transitioned from standalone systems into essential components across broader applications, with indexing efficiency, cost-effectiveness, and freshness becoming increasingly critical yet often overlooked. In this paper, we introduce Seml-parametric Disentangled Retrieval (SiDR), a bi-encoder retrieval framework that decouples retrieval index from neural parameters to enable efficient, low-cost, and parameter-agnostic indexing for emerging use cases. Specifically, in addition to using embeddings as indexes like existing neural retrieval methods, SiDR supports a non-parametric tokenization index for search, achieving BM25-like indexing complexity with significantly better effectiveness. Our comprehensive evaluation across 16 retrieval benchmarks demonstrates that SiDR outperforms both neural and term-based retrieval baselines under the same indexing workload: (i) When using an embedding-based index, SiDR exceeds the performance of conventional neural retrievers while maintaining similar training complexity; (ii) When using a tokenization-based index, SiDR drastically reduces indexing cost and time, matching the complexity of traditional term-based retrieval, while consistently outperforming BM25 on all in-domain datasets; (iii) Additionally, we introduce a late parametric mechanism that matches BM25 index preparation time while outperforming other neural retrieval baselines in effectiveness.

1547. M^3PC: Test-time Model Predictive Control using Pretrained Masked Trajectory Model

链接：https://iclr.cc/virtual/2025/poster/28673 abstract： Recent work in Offline Reinforcement Learning (RL) has shown that a unified transformer trained under a masked auto-encoding objective can effectively capture the relationships between different modalities (e.g., states, actions, rewards) within given trajectory datasets. However, this information has not been fully exploited during the inference phase, where the agent needs to generate an optimal policy instead of just reconstructing masked components from unmasked. Given that a pretrained trajectory model can act as both a Policy Model and a World Model with appropriate mask patterns, we propose using Model Predictive Control (MPC) at test time to leverage the model's own predictive capacity to guide its action selection. Empirical results on D4RL and RoboMimic show that our inference-phase MPC significantly improves the decision-making performance of a pretrained trajectory model without any additional parameter training. Furthermore, our framework can be adapted to Offline to Online (O2O) RL and Goal Reaching RL, resulting in more substantial performance gains when an additional online interaction budget is given, and better generalization capabilities when different task targets are specified. Code is available: \href{https://github.com/wkh923/m3pc} {\texttt{https://github.com/wkh923/m3pc}}.

1548. Imputation for prediction: beware of diminishing returns.

链接：https://iclr.cc/virtual/2025/poster/30481 abstract： Missing values are prevalent across various fields, posing challenges for training and deploying predictive models. In this context, imputation is a common practice, driven by the hope that accurate imputations will enhance predictions. However, recent theoretical and empirical studies indicate that simple constant imputation can be consistent and competitive. This empirical study aims at clarifying if and when investing in advanced imputation methods yields significantly better predictions. Relating imputation and predictive accuracies across combinations of imputation and predictive models on 19 datasets, we show that imputation accuracy matters less i) when using expressive models, ii) when incorporating missingness indicators as complementary inputs, iii) matters much more for generated linear outcomes than for

real-data outcomes. Interestingly, we also show that the use of the missingness indicator is beneficial to the prediction performance, even in MCAR scenarios. Overall, on real-data with powerful models, imputation quality has only a minor effect on prediction performance. Thus, investing in better imputations for improved predictions often offers limited benefits.

1549. Forking Paths in Neural Text Generation

链接：https://iclr.cc/virtual/2025/poster/30769 abstract： Estimating uncertainty in Large Language Models (LLMs) is important for properly evaluating LLMs, and ensuring safety for users. However, prior approaches to uncertainty estimation focus on the final answer in generated text, ignoring intermediate steps that might dramatically impact the outcome. We hypothesize that there exist key forking tokens, such that re-sampling the system at those specific tokens, but not others, leads to very different outcomes. To test this empirically, we develop a novel approach to representing uncertainty dynamics across individual tokens of text generation, and applying statistical models to test our hypothesis. Our approach is highly flexible: it can be applied to any dataset and any LLM, without fine tuning or accessing model weights. We use our method to analyze LLM responses on 7 different tasks across 4 domains, spanning a wide range of typical use cases. We find many examples of forking tokens, including surprising ones such as a space character instead of a colon, suggesting that LLMs are often just a single token away from saying something very different.

1550. RECAST: Reparameterized, Compact weight Adaptation for Sequential Tasks

链接：https://iclr.cc/virtual/2025/poster/30133 abstract： Incremental learning aims to adapt to new sets of categories over time with minimal computational overhead. Prior work often addresses this task by training efficient task-specific adaptors that modify frozen layer weights or features to capture relevant information without affecting predictions on any previously learned categories. While these adaptors are generally more efficient than finetuning the entire network, they still can require tens to hundreds of thousands task-specific trainable parameters even for relatively small networks. This makes it challenging to operate in resource-constrained environments with high communication costs, such as edge devices or mobile phones. Thus, we propose Reparameterized, Compact weight Adaptation for Sequential Tasks (RECAST), a novel method that dramatically reduces the number of task-specific trainable parameters to fewer than 50 – several orders of magnitude less than competing methods like LoRA. RECAST accomplishes this efficiency by learning to decompose layer weights into a soft parameter-sharing framework consisting of a set of shared weight templates and very few module-specific scaling factors or coefficients. This soft parameter-sharing framework allows for effective task-wise reparameterization by tuning only these coefficients while keeping templates frozen. A key innovation of RECAST is the novel weight reconstruction pipeline called Neural Mimicry, which eliminates the need for pretraining from scratch. This allows for high-fidelity emulation of existing pretrained weights within our framework and provides quick adaptability to any model scale and architecture. Extensive experiments across six diverse datasets demonstrate RECAST outperforms the state-of-the-art by up to ~ 1.5% and improves baselines by > 3% across various scales, architectures, and parameter spaces. Moreover, we show that RECAST's architecture-agnostic nature allows for seamless integration with existing methods, further boosting performance.

1551. Broaden your SCOPE! Efficient Multi-turn Conversation Planning for LLMs with Semantic Space

链接：https://iclr.cc/virtual/2025/poster/31066 abstract： Large language models (LLMs) are used in chatbots or AI assistants to hold conversations with a human user. In such applications, the quality (e.g., user engagement, safety) of a conversation is important and can only be exactly known at the end of the conversation. To maximize its expected quality, conversation planning reasons about the stochastic transitions within a conversation to select the optimal LLM response at each turn. Existing simulation-based conversation planning algorithms typically select the optimal response by simulating future conversations with a large number of LLM queries at every turn. However, this process is extremely time-consuming and hence impractical for real-time conversations. This paper presents a novel approach called Semantic space COnversation Planning with improved Efficiency (SCOPE) that exploits the dense semantic representation of conversations to perform conversation planning efficiently. In particular, SCOPE models the stochastic transitions in conversation semantics and their associated rewards to plan entirely within the semantic space. This allows us to select the optimal LLM response at every conversation turn without needing additional LLM queries for simulation. As a result, SCOPE can perform conversation planning 70 times faster than conventional simulation-based planning algorithms when applied to a wide variety of conversation starters and two reward functions seen in the real world, yet achieving a higher reward within a practical planning budget. Our code can be found at: https://github.com/chenzhiliang94/convo-plan-SCOPE.

1552. HALL-E: Hierarchical Neural Codec Language Model for Minute-Long Zero-Shot Text-to-Speech Synthesis

链接：https://iclr.cc/virtual/2025/poster/30784 abstract： Recently, Text-to-speech (TTS) models based on large language models (LLMs)that translate natural language text into sequences of discrete audio tokens havegained great research attention, with advances in neural audio codec (NAC) mod-els using residual vector quantization (RVQ). However, long-form speech synthe-sis remains a significant challenge due to the high frame rate, which increases thelength of audio tokens and makes it difficult for autoregressive language modelsto generate audio tokens for even a minute of speech. To address this challenge,this

paper introduces two novel post-training approaches: 1) Multi-Resolution Re-quantization (MReQ) and 2) HALL-E. MReQ is a framework to reduce the framerate of pre-trained NAC models. Specifically, it incorporates multi-resolutionresidual vector quantization (MRVQ) module that hierarchically reorganizes dis-crete audio tokens through teacher-student distillation. HALL-E is an LLM-basedTTS model designed to predict hierarchical tokens of MReQ. Specifically, it incor-porates the technique of using MRVQ sub-modules and continues training from apre-trained LLM-based TTS model. Furthermore, to promote TTS research, wecreate MinutesSpeech, a new benchmark dataset consisting of 40k hours of filteredspeech data for training and evaluating speech synthesis ranging from 3s up to180s. In experiments, we demonstrated the effectiveness of our approaches by ap-plying our post-training framework to VALL-E. We achieved the frame rate downto as low as 8 Hz, enabling the stable minitue-long speech synthesis in a singleinference step. Audio samples, dataset, codes and pre-trained models are availableat https://yutonishimura-v2.github.io/HALL-E_DEMO.

1553. Adapters for Altering LLM Vocabularies: What Languages Benefit the Most?

链接：https://iclr.cc/virtual/2025/poster/30024 abstract： Vocabulary adaptation, which integrates new vocabulary into pre-trained language models, enables expansion to new languages and mitigates token over-fragmentation. However, existing approaches are limited by their reliance on heuristics or external embeddings. We propose VocADT, a novel method for vocabulary adaptation using adapter modules that are trained to learn the optimal linear combination of existing embeddings while keeping the model’s weights fixed. VocADT offers a flexible and scalable solution without depending on external resources or language constraints. Across 11 languages—with diverse scripts, resource availability, and fragmentation—we demonstrate that VocADT outperforms the original Mistral model (Jiang et al., 2023) and other baselines across various multilingual tasks including natural language understanding and machine translation. We find that Latin-script languages and highly fragmented languagesbenefit the most from vocabulary adaptation. We further fine-tune the adapted model on the generative task of machine translation and find that vocabulary adaptation is still beneficial after fine-tuning and that VocADT is the most effective.

1554. Scaling LLM Test-Time Compute Optimally Can be More Effective than Scaling Parameters for Reasoning

链接：https://iclr.cc/virtual/2025/poster/31024 abstract： Enabling LLMs to improve their outputs by using more test-time compute is a critical step towards building self-improving agents that can operate on open-ended natural language. In this paper, we scale up inference-time computation in LLMs, with a focus on answering: if an LLM is allowed to use a fixed but non-trivial amount of inference-time compute, how much can it improve its performance on a challenging prompt? Answering this question has implications not only on performance, but also on the future of LLM pretraining and how to tradeoff inference-time and pre-training compute. Little research has attempted to understand the scaling behaviors of test-time inference methods, with current work largely providing negative results for a number of these strategies. In this work, we analyze two primary mechanisms to scale test-time computation: (1) searching against dense, process-based verifier reward models (PRMs); and (2) updating the model's distribution over a response adaptively, given the prompt at test time. We find that in both cases, the effectiveness of different approaches to scaling test-time compute critically varies depending on the difficulty of the prompt. This observation motivates applying a "compute-optimal" scaling strategy, which acts to, as effectively as possible, allocate test-time compute per prompt in an adaptive manner. Using this compute-optimal strategy, we can improve the efficiency of test-time compute scaling for math reasoning problems by more than 4x compared to a best-of-N baseline. Additionally, in a FLOPs-matched evaluation, we find that on problems where a smaller base model attains somewhat non-trivial success rates, test-time compute can be used to outperform a 14x larger model.

1555. Model Equality Testing: Which Model is this API Serving?

链接：https://iclr.cc/virtual/2025/poster/29704 abstract： Users often interact with large language models through black-box inference APIs, both for closed- and open-weight models (e.g., Llama models are popularly accessed via Amazon Bedrock and Azure AI Studio). In order to cut costs or add functionality, API providers may quantize, watermark, or finetune the underlying model, changing the output distribution --- possibly without notifying users. We formalize detecting such distortions as Model Equality Testing, a two-sample testing problem, where the user collects samples from the API and a reference distribution and conducts a statistical test to see if the two distributions are the same. We find that tests based on the Maximum Mean Discrepancy between distributions are powerful for this task: a test built on a simple string kernel achieves a median of 77.4% power against a range of distortions, using an average of just 10 samples per prompt. We then apply this test to commercial inference APIs from Summer 2024 for four Llama models, finding that 11 out of 31 endpoints serve different distributions than reference weights released by Meta.

1556. Decoupled Finetuning for Domain Generalizable Semantic Segmentation

链接：https://iclr.cc/virtual/2025/poster/28246 abstract： Joint finetuning of a pretrained encoder and a randomly initialized decoder has been the de facto standard in semantic segmentation, but the vulnerability of this approach to domain shift has not been studied. We investigate the vulnerability issue of joint finetuning, and propose a novel finetuning framework called Decoupled FineTuning (DeFT) for domain generalization as a solution. DeFT operates in two stages. Its first stage warms up

the decoder with the frozen, pretrained encoder so that the decoder learns task-relevant knowledge while the encoder preserves its generalizable features. In the second stage, it decouples finetuning of the encoder and decoder into two pathways, each of which concatenates an adaptive component (AC) and retentive component (RC); the encoder and decoder play different roles between AC and RC in different pathways. ACs are updated by gradients of the loss on the source domain, while RCs are updated by exponential moving average biased toward their initialization to retain their generalization capability. By the two separate optimization pathways with opposite AC-RC configurations, DeFT reduces the number of learnable parameters virtually, and decreases the distance between learned parameters and their initialization, leading to improved generalization capability. DeFT significantly outperformed existing methods in various domain shift scenarios, and its performance was further boosted by incorporating a simple distance regularization.

1557. Interpretable Vision-Language Survival Analysis with Ordinal Inductive Bias for Computational Pathology

链接：https://iclr.cc/virtual/2025/poster/28016 abstract：

1558. Context Clues: Evaluating Long Context Models for Clinical Prediction Tasks on EHR Data

链接：https://iclr.cc/virtual/2025/poster/27657 abstract： Foundation Models (FMs) trained on Electronic Health Records (EHRs) have achieved state-of-the-art results on numerous clinical prediction tasks. However, prior EHR FMs typically have context windows of $<$1k tokens, which prevents them from modeling full patient EHRs which can exceed 10k's of events. For making clinical predictions, both model performance and robustness to the unique properties of EHR data are crucial. Recent advancements in subquadratic long-context architectures (e.g. Mamba) offer a promising solution. However, their application to EHR data has not been well-studied. We address this gap by presenting the first systematic evaluation of the effect of context length on modeling EHR data. We find that longer context models improve predictive performance -- our Mamba-based model surpasses the prior state-of-the-art on 9/14 tasks on the EHRSHOT prediction benchmark. Additionally, we measure robustness to three unique, previously underexplored properties of EHR data: (1) the prevalence of ``copy-forwarded'' diagnoses which create artificial token repetition in EHR sequences; (2) the irregular time intervals between EHR events which can lead to a wide range of timespans within a context window; and (3) the natural increase in disease complexity over time which makes later tokens in the EHR harder to predict than earlier ones. Stratifying our EHRSHOT results, we find that higher levels of each property correlate negatively with model performance (e.g., a 14% higher Brier loss between the least and most irregular patients), but that longer context models are more robust to more extreme levels of these properties. Our work highlights the potential for using long-context architectures to model EHR data, and offers a case study on how to identify and quantify new challenges in modeling sequential data motivated by domains outside of natural language. We release all of our model checkpoints and code.

1559. Enhancing Uncertainty Estimation and Interpretability with Bayesian Non-negative Decision Layer

链接：https://iclr.cc/virtual/2025/poster/27792 abstract： Although deep neural networks have demonstrated significant success due to theirpowerful expressiveness, most models struggle to meet practical requirements foruncertainty estimation. Concurrently, the entangled nature of deep neural net-works leads to a multifaceted problem, where various localized explanation tech-niques reveal that multiple unrelated features influence the decisions, thereby un-dermining interpretability. To address these challenges, we develop a BayesianNonnegative Decision Layer (BNDL), which reformulates deep neural networksas a conditional Bayesian non-negative factor analysis. By leveraging stochasticlatent variables, the BNDL can model complex dependencies and provide robustuncertainty estimation. Moreover, the sparsity and non-negativity of the latentvariables encourage the model to learn disentangled representations and decisionlayers, thereby improving interpretability. We also offer theoretical guaranteesthat BNDL can achieve effective disentangled learning. In addition, we developeda corresponding variational inference method utilizing a Weibull variational in-ference network to approximate the posterior distribution of the latent variables.Our experimental results demonstrate that with enhanced disentanglement capa-bilities, BNDL not only improves the model’s accuracy but also provides reliableuncertainty estimation and improved interpretability.

1560. Progressive Token Length Scaling in Transformer Encoders for Efficient Universal Segmentation

链接：https://iclr.cc/virtual/2025/poster/28965 abstract： A powerful architecture for universal segmentation relies on transformers that encode multi-scale image features and decode object queries into mask predictions. With efficiency being a high priority for scaling such models, we observed that the state-of-the-art method Mask2Former uses \~50% of its compute only on the transformer encoder. This is due to the retention of a full-length token-level representation of all backbone feature scales at each encoder layer. With this observation, we propose a strategy termed PROgressive Token Length SCALing for Efficient transformer encoders (PRO-SCALE) that can be plugged-in to the Mask2Former segmentation architecture to significantly reduce the computational cost. The underlying principle of PRO-SCALE is: progressively scale the length of the tokens with the layers of the encoder. This allows PRO-SCALE to reduce computations by a large margin with minimal sacrifice in performance (\~52% encoder and \~27% overall GFLOPs reduction with no drop in performance on COCO dataset). Experiments conducted

on public benchmarks demonstrates PRO-SCALE's flexibility in architectural configurations, and exhibits potential for extension beyond the settings of segmentation tasks to encompass object detection. Code is available here: https://github.com/abhishekaich27/proscale-pytorch

1561. Toward Guidance-Free AR Visual Generation via Condition Contrastive Alignment

链接：https://iclr.cc/virtual/2025/poster/28596 abstract： Classifier-Free Guidance (CFG) is a critical technique for enhancing the sample quality of visual generative models. However, in autoregressive (AR) multi-modal generation, CFG introduces design inconsistencies between language and visual content, contradicting the design philosophy of unifying different modalities for visual AR. Motivated by language model alignment methods, we propose Condition Contrastive Alignment (CCA) to facilitate guidance-free AR visual generation. Unlike guidance methods that alter the sampling process to achieve the ideal sampling distribution, CCA directly fine-tunes pretrained models to fit the same distribution target. Experimental results show that CCA can significantly enhance the guidance-free performance of all tested models with just one epoch of fine-tuning (1% of pretraining epochs) on the pretraining dataset. This largely removes the need for guided sampling in AR visual generation and cuts the sampling cost by half. Moreover, by adjusting training parameters, CCA can achieve trade-offs between sample diversity and fidelity similar to CFG. This experimentally confirms the strong theoretical connection between language-targeted alignment and visual-targeted guidance methods, unifying two previously independent research fields.

1562. Positive-Unlabeled Diffusion Models for Preventing Sensitive Data Generation

链接：https://iclr.cc/virtual/2025/poster/28640 abstract： Diffusion models are powerful generative models but often generate sensitive data that are unwanted by users,mainly because the unlabeled training data frequently contain such sensitive data.Since labeling all sensitive data in the large-scale unlabeled training data is impractical,we address this problem by using a small amount of labeled sensitive data.In this paper,we propose positive-unlabeled diffusion models,which prevent the generation of sensitive data using unlabeled and sensitive data.Our approach can approximate the evidence lower bound (ELBO) for normal (negative) data using only unlabeled and sensitive (positive) data.Therefore, even without labeled normal data,we can maximize the ELBO for normal data and minimize it for labeled sensitive data,ensuring the generation of only normal data.Through experiments across various datasets and settings,we demonstrated that our approach can prevent the generation of sensitive images without compromising image quality.

1563. Equivariant Neural Functional Networks for Transformers

链接：https://iclr.cc/virtual/2025/poster/27995 abstract： This paper systematically explores neural functional networks (NFN) for transformer architectures. NFN are specialized neural networks that treat the weights, gradients, or sparsity patterns of a deep neural network (DNN) as input data and have proven valuable for tasks such as learnable optimizers, implicit data representations, and weight editing. While NFN have been extensively developed for MLP and CNN, no prior work has addressed their design for transformers, despite the importance of transformers in modern deep learning. This paper aims to address this gap by providing a systematic study of NFN for transformers. We first determine the maximal symmetric group of the weights in a multi-head attention module as well as a necessary and sufficient condition under which two sets of hyperparameters of the multi-head attention module define the same function. We then define the weight space of transformer architectures and its associated group action, which leads to the design principles for NFN in transformers. Based on these, we introduce Transformer-NFN, an NFN that is equivariant under this group action. Additionally, we release a dataset of more than 125,000 Transformers model checkpoints trained on two datasets with two different tasks, providing a benchmark for evaluating Transformer-NFN and encouraging further research on transformer training and performance.

1564. Intermediate Layer Classifiers for OOD generalization

链接：https://iclr.cc/virtual/2025/poster/30537 abstract： Deep classifiers are known to be sensitive to data distribution shifts, primarily due to their reliance on spurious correlations in training data. It has been suggested that these classifiers can still find useful features in the network's last layer that hold up under such shifts. In this work, we question the use of last-layer representations for out-of-distribution (OOD) generalisation and explore the utility of intermediate layers. To this end, we introduce \textit{Intermediate Layer Classifiers} (ILCs). We discover that intermediate layer representations frequently offer substantially better generalisation than those from the penultimate layer. In many cases, zero-shot OOD generalisation using earlier-layer representations approaches the few-shot performance of retraining on penultimate layer representations. This is confirmed across multiple datasets, architectures, and types of distribution shifts. Our analysis suggests that intermediate layers are less sensitive to distribution shifts compared to the penultimate layer. These findings highlight the importance of understanding how information is distributed across network layers and its role in OOD generalisation, while also pointing to the limits of penultimate layer representation utility. Code is available at https://github.com/oshapio/intermediate-layer-generalization.

1565. Jamba: Hybrid Transformer-Mamba Language Models

链接：https://iclr.cc/virtual/2025/poster/30115 abstract： We present Jamba, a novel hybrid Transformer-Mamba mixture-of-experts (MoE) architecture. Jamba interleaves blocks of Transformer and Mamba layers, enjoying the benefits of both model families. MoE is added in some of these layers to increase model capacity while keeping active parameter usage manageable. This flexible architecture allows resource- and objective-specific configurations. We implement two configurations: Jamba-1.5-Large, with 94B active parameters, and Jamba-1.5-mini, with 12B active parameters. Built at large scale, Jamba models provide high throughput and small memory footprint compared to vanilla Transformers, especially at long-context tasks, with an effective context length of 256K tokens, the largest amongst open-weight models. At the same time, they are also competitive on standard language modeling and chatbot benchmarks. We study various architectural decisions, such as how to combine Transformer and Mamba layers, and how to mix experts, and show that some of them are crucial in large scale modeling. To support cost-effective inference, we introduce ExpertsInt8, a novel quantization technique that allows fitting Jamba-1.5-Large on a machine with 8 80GB GPUs when processing 256K-token contexts without loss of quality. We also describe several interesting properties of this architecture that the training and evaluation of Jamba have revealed. The model weights are publicly available.

1566. From Tokens to Words: On the Inner Lexicon of LLMs

链接：https://iclr.cc/virtual/2025/poster/31096 abstract： Natural language is composed of words, but modern large language models (LLMs) process sub-words as input. A natural question raised by this discrepancy is whether LLMs encode words internally, and if so how. We present evidence that LLMs engage in an intrinsic detokenization process, where subword sequences are combined into coherent whole-word representations at their last token. Our experiments show that this process primarily takes place within the early and middle layers of the model. We further demonstrate its robustness to arbitrary splits (e.g., “cats” to “ca” and “ts”), typos, and importantly—to out-of-vocabulary words: when feeding the last token internal representations of such words to the model as input, it can “understand” them as the complete word despite never seeing such representations as input during training. Our findings suggest that LLMs maintain a latent vocabulary beyond the tokenizer’s scope. These insights provide a practical, finetuning-free application for expanding the vocabulary of pre-trained models. By enabling the addition of new vocabulary words, we reduce input length and inference iterations, which reduces both space and model latency, with little to no loss in model accuracy.

1567. A Curious Case of the Missing Measure: Better Scores and Worse Generation

链接：https://iclr.cc/virtual/2025/poster/31366 abstract： Our field has a secret: nobody fully trusts audio evaluation measures. As neural audio generation nears perceptual fidelity, these measures fail to detect subtle differences that human listeners readily identify, often contradicting each other when comparing state-of-the-art models. The gap between human perception and automatic measures means we have increasingly sophisticated models while losing our ability to understand their flaws.

1568. Add-it: Training-Free Object Insertion in Images With Pretrained Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29207 abstract： Adding Object into images based on text instructions is a challenging task in semantic image editing, requiring a balance between preserving the original scene and seamlessly integrating the new object in a fitting location. Despite extensive efforts, existing models often struggle with this balance, particularly with finding a natural location for adding an object in complex scenes. We introduce Add-it, a training-free approach that extends diffusion models' attention mechanisms to incorporate information from three key sources: the scene image, the text prompt, and the generated image itself. Our weighted extended-attention mechanism maintains structural consistency and fine details while ensuring natural object placement. Without task-specific fine-tuning, Add-it achieves state-of-the-art results on both real and generated image insertion benchmarks, including our newly constructed "Additing Affordance Benchmark" for evaluating object placement plausibility, outperforming supervised methods. Human evaluations show that Add-it is preferred in over 80% of cases, and it also demonstrates improvements in various automated metrics.

1569. Self-Updatable Large Language Models by Integrating Context into Model Parameters

链接：https://iclr.cc/virtual/2025/poster/29183 abstract： Despite significant advancements in large language models (LLMs), the rapid and frequent integration of small-scale experiences, such as interactions with sur- rounding objects, remains a substantial challenge. Two critical factors in assimilating these experiences are (1) Efficacy: the ability to accurately remember recent events; (2) Retention: the capacity to recall long-past experiences. Current methods either embed experiences within model parameters using continual learning, model editing, or knowledge distillation techniques, which often struggle with rapid updates and complex interactions, or rely on external storage to achieve long-term retention, thereby increasing storage requirements. In this paper, we propose SELF-PARAM (Self-Updatable Large Language Models with Parameter Integration). SELF-PARAM requires no extra parameters while ensuring near-optimal efficacy and long-term retention. Our method employs a training objective that minimizes the Kullback-Leibler (KL) divergence between the predictions of an original model (with access to contextual information) and a target model (without such access). By generating diverse question-answer pairs related to the knowledge and minimizing the KL divergence across this dataset, we update the target model to internalize the

knowledge seamlessly within its parameters. Evaluations on question-answering and conversational recommendation tasks demonstrate that SELF-PARAM significantly outperforms existing methods, even when accounting for non-zero storage requirements. This advancement paves the way for more efficient and scalable integration of experiences in large language models by embedding knowledge directly into model parameters.

1570. Exploiting Distribution Constraints for Scalable and Efficient Image Retrieval

链接：https://iclr.cc/virtual/2025/poster/29018 abstract： Image retrieval is crucial in robotics and computer vision, with downstream applications in robot place recognition and vision-based product recommendations. Modern retrieval systems face two key challenges: scalability and efficiency.State-of-the-art image retrieval systems train specific neural networks for each dataset, an approach that lacks scalability. Furthermore, since retrieval speed is directly proportional to embedding size, existing systems that use large embeddings lack efficiency. To tackle scalability, recent works propose using off-the-shelf foundation models. However, these models, though applicable across datasets, fall short in achieving performance comparable to that of dataset-specific models. Our key observation is that, while foundation models capture necessary subtleties for effective retrieval, the underlying distribution of their embedding space can negatively impact cosine similarity searches. We introduce Autoencoders with Strong Variance Constraints (AE-SVC), which, when used for projection, significantly improves the performance of foundation models. We provide an in-depth theoretical analysis of AE-SVC. Addressing efficiency, we introduce Single-Shot Similarity Space Distillation ((SS)2D), a novel approach to learn embeddings with adaptive sizes that offers a better trade-off between size and performance. We conducted extensive experiments on four retrieval datasets, including Stan-ford Online Products (SoP) and Pittsburgh30k, using four different off-the-shelf foundation models, including DinoV2 and CLIP. AE-SVC demonstrates up to a 16% improvement in retrieval performance, while (SS)2D shows a further 10% improvement for smaller embedding sizes.

1571. Dimension Agnostic Neural Processes

链接：https://iclr.cc/virtual/2025/poster/27989 abstract： Meta-learning aims to train models that can generalize to new tasks with limited labeled data by extracting shared features across diverse task datasets. Additionally, it accounts for prediction uncertainty during both training and evaluation, a concept known as uncertainty-aware meta-learning. Neural Process (NP) is a well-known uncertainty-aware meta-learning method that constructs implicit stochastic processes using parametric neural networks, enabling rapid adaptation to new tasks. However, existing NP methods face challenges in accommodating diverse input dimensions and learned features, limiting their broad applicability across regression tasks. To address these limitations and advance the utility of NP models as general regressors, we introduce Dimension Agnostic Neural Process (DANP). DANP incorporates Dimension Aggregator Block (DAB) to transform input features into a fixed-dimensional space, enhancing the model's ability to handle diverse datasets. Furthermore, leveraging the Transformer architecture and latent encoding layers, DANP learns a wider range of features that are generalizable across various tasks. Through comprehensive experimentation on various synthetic and practical regression tasks, we empirically show that DANP outperforms previous NP variations, showcasing its effectiveness in overcoming the limitations of traditional NP models and its potential for broader applicability in diverse regression scenarios.

1572. HR-Extreme: A High-Resolution Dataset for Extreme Weather Forecasting

链接：https://iclr.cc/virtual/2025/poster/30964 abstract： The application of large deep learning models in weather forecasting has led tosignificant advancements in the field, including higher-resolution forecasting andextended prediction periods exemplified by models such as Pangu and Fuxi. Despitethese successes, previous research has largely been characterized by the neglectof extreme weather events, and the availability of datasets specifically curated forsuch events remains limited. Given the critical importance of accurately forecastingextreme weather, this study introduces a comprehensive dataset that incorporateshigh-resolution extreme weather cases derived from the High-Resolution RapidRefresh (HRRR) data, a 3-km real-time dataset provided by NOAA. We alsoevaluate the current state-of-the-art deep learning models and Numerical WeatherPrediction (NWP) systems on HR-Extreme, and provide a improved baselinedeep learning model called HR-Heim which has superior performance on bothgeneral loss and HR-Extreme compared to others. Our results reveal that theerrors of extreme weather cases are significantly larger than overall forecast error,highlighting them as an crucial source of loss in weather prediction. These findingsunderscore the necessity for future research to focus on improving the accuracy ofextreme weather forecasts to enhance their practical utility

1573. ViSAGe: Video-to-Spatial Audio Generation

链接：https://iclr.cc/virtual/2025/poster/30759 abstract： Spatial audio is essential for enhancing the immersiveness of audio-visual experiences, yet its production typically demands complex recording systems and specialized expertise. In this work, we address a novel problem of generating first-order ambisonics, a widely used spatial audio format, directly from silent videos. To support this task, we introduce YT-Ambigen, a dataset comprising 102K 5-second YouTube video clips paired with corresponding first-order ambisonics. We also propose new evaluation metrics to assess the spatial aspect of generated audio based on audio energy maps and saliency metrics. Furthermore, we present Video-to-Spatial Audio Generation (ViSAGe), an end-to-end framework that generates first-order ambisonics from silent video frames by leveraging CLIP visual features,

autoregressive neural audio codec modeling with both directional and visual guidance. Experimental results demonstrate that ViSAGe produces plausible and coherent first-order ambisonics, outperforming two-stage approaches consisting of video-to-audio generation and audio spatialization. Qualitative examples further illustrate that ViSAGe generates temporally aligned high-quality spatial audio that adapts to viewpoint changes.

1574. Revisiting Energy Based Models as Policies: Ranking Noise Contrastive Estimation and Interpolating Energy Models

链接：https://iclr.cc/virtual/2025/poster/31478 abstract： A crucial design decision for any robot learning pipeline is the choice of policy representation: what type of model should be used to generate the next set of robot actions? Owing to the inherent multi-modal nature of many robotic tasks, combined with the recent successes in generative modeling, researchers have turned to state-of-the-art probabilistic models such as diffusion models for policy representation. In this work, we revisit the choice of energy-based models (EBM) as a policy class. We show that the prevailing folklore---that energy models in high dimensional continuous spaces are impractical to train---is false. We develop a practical training objective and algorithm for energy models which combines several key ingredients: (i) ranking noise contrastive estimation (R-NCE), (ii) learnable negative samplers, and (iii) non-adversarial joint training. We prove that our proposed objective function is asymptotically consistent and quantify its limiting variance. On the other hand, we show that the Implicit Behavior Cloning (IBC) objective is actually biased even at the population level, providing a mathematical explanation for the poor performance of IBC trained energy policies in several independent follow-up works. We further extend our algorithm to learn a continuous stochastic process that bridges noise and data, modeling this process with a family of EBMs indexed by scale variable. In doing so, we demonstrate that the core idea behind recent progress in generative modeling is actually compatible with EBMs. Altogether, our proposed training algorithms enable us to train energy-based models as policies which compete with---and even outperform---diffusion models and other state-of-the-art approaches in several challenging multi-modal benchmarks: obstacle avoidance path planning and contact-rich block pushing.

1575. Fine-tuning with Reserved Majority for Noise Reduction

链接：https://iclr.cc/virtual/2025/poster/29216 abstract： Parameter-efficient fine-tuning (PEFT) has revolutionized supervised fine-tuning, where LoRA and its variants gain the most popularity due to their low training costs and zero inference latency.However, LoRA tuning not only injects knowledgeable features but also noisy hallucination during fine-tuning, which hinders the utilization of tunable parameters with the increasing LoRA rank.In this work, we first investigate in-depth the redundancies among LoRA parameters with substantial empirical studies.Aiming to resemble the learning capacity of high ranks from the findings, we set up a new fine-tuning framework, \textbf{P}arameter-\textbf{Re}dundant \textbf{F}ine-\textbf{T}uning (\preft), which follows the vanilla LoRA tuning process but is required to reduce redundancies before merging LoRA parameters back to pre-trained models.Based on this framework, we propose \textbf{No}ise reduction with \textbf{R}eserved \textbf{M}ajority~(\norm), which decomposes the LoRA parameters into majority parts and redundant parts with random singular value decomposition.The major components are determined by the proposed \search method, specifically employing subspace similarity to confirm the parameter groups that share the highest similarity with the base weight.By employing \norm, we enhance both the learning capacity and benefits from larger ranks, which consistently outperforms both LoRA and other \preft-based methods on various downstream tasks, such as general instruction tuning, math reasoning and code generation. Code is available at \url{https://github.com/pixas/NoRM}.

1576. Shallow diffusion networks provably learn hidden low-dimensional structure

链接：https://iclr.cc/virtual/2025/poster/30033 abstract： Diffusion-based generative models provide a powerful framework for learning to sample from a complex target distribution. The remarkable empirical success of these models applied to high-dimensional signals, including images and video, stands in stark contrast to classical results highlighting the curse of dimensionality for distribution recovery. In this work, we take a step towards understanding this gap through a careful analysis of learning diffusion models over the Barron space of single hidden layer neural networks. In particular, we show that these shallow models provably adapt to simple forms of low-dimensional structure, such as an unknown linear subspace or hidden independence, thereby avoiding the curse of dimensionality. We combine our results with recent analyses of sampling with diffusions to provide an end-to-end sample complexity bound for learning to sample from structured distributions. Importantly, our results do not require specialized architectures tailored to particular latent structures, and instead rely on the low-index structure of the Barron space to adapt to the underlying distribution.

1577. Deep Kernel Posterior Learning under Infinite Variance Prior Weights

链接：https://iclr.cc/virtual/2025/poster/27952 abstract： Neal (1996) proved that infinitely wide shallow Bayesian neural networks (BNN) converge to Gaussian processes (GP), when the network weights have bounded prior variance. Cho & Saul (2009) provided a useful recursive formula for deep kernel processes for relating the covariance kernel of each layer to the layer immediately below. Moreover, they worked out the form of the layer-wise covariance kernel in an explicit manner for several common activation functions, including the ReLU. Subsequent works have made the connection between these two works, and provided useful results on the covariance kernel of a deep GP arising as wide limits of various deep Bayesian network architectures. However, recent works, including Aitchison et al. (2021), have highlighted that the covariance kernels obtained in

this manner are deterministic and hence, precludes any possibility of representation learning, which amounts to learning a non-degenerate posterior of a random kernel given the data. To address this, they propose adding artificial noise to the kernel to retain stochasticity, and develop deep kernel Wishart and inverse Wishart processes. Nonetheless, this artificial noise injection could be critiqued in that it would not naturally emerge in a classic BNN architecture under an infinite-width limit. To address this, we show that a Bayesian deep neural network, where each layer width approaches infinity, and all network weights are elliptically distributed with infinite variance, converges to a process with α-stable marginals in each layer that has a conditionally Gaussian representation. These conditional random covariance kernels could be recursively linked in the manner of Cho & Saul (2009), even though marginally the process exhibits stable behavior, and hence covariances are not even necessarily defined. We also provide useful generalizations of the recent results of Loría & Bhadra (2024) on shallow networks to multi-layer networks, and remedy the prohibitive computational burden of their approach. The computational and statistical benefits over competing approaches stand out in simulations and in demonstrations on benchmark data sets.

1578. Credit-based self organizing maps: training deep topographic networks with minimal performance degradation

链接：https://iclr.cc/virtual/2025/poster/27851 abstract：

1579. Privacy Auditing of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30914 abstract： Current techniques for privacy auditing of large language models (LLMs) have limited efficacy---they rely on basic approaches to generate canaries which leads to weak membership inference attacks that in turn give loose lower bounds on the empirical privacy leakage.We develop canaries that are far more effective than those used in prior work under threat models that cover a range of realistic settings. We demonstrate through extensive experiments on multiple families of fine-tuned LLMs that our approach sets a new standard for detection of privacy leakage. For measuring the memorization rate of non-privately trained LLMs, our designed canaries surpass prior approaches. For example, on the Qwen2.5-0.5B model, our designed canaries achieve 49.6% TPR at 1% FPR, vastly surpassing the prior approach's 4.2% TPR at 1% FPR. Our method can be used to provide a privacy audit of $\varepsilon \approx 1$ for a model trained with theoretical ε of 4. To the best of our knowledge, this is the first time that a privacy audit of LLM training has achieved nontrivial auditing success in the setting where the attacker cannot train shadow models, insert gradient canaries, or access the model at every iteration.

1580. Distilled Decoding 1: One-step Sampling of Image Auto-regressive Models with Flow Matching

链接：https://iclr.cc/virtual/2025/poster/27673 abstract： Autoregressive (AR) models have recently achieved state-of-the-art performance in text and image generation. However, their primary limitation is slow generation speed due to the token-by-token process. We ask an ambitious question: can a pre-trained AR model be adapted to generate outputs in just one or two steps? If successful, this would significantly advance the development and deployment of AR models. We notice that existing works that attempt to speed up AR generation by generating multiple tokens at once fundamentally cannot capture the output distribution due to the conditional dependencies between tokens, limiting their effectiveness for few-step generation. To overcome this, we propose Distilled Decoding (DD), which leverages flow matching to create a deterministic mapping from Gaussian distribution to the output distribution of the pre-trained AR model. We then train a network to distill this mapping, enabling few-step generation. The entire training process of DD does not need the training data of the original AR model (as opposed to some other methods), thus making DD more practical. We evaluate DD on state-of-the-art image AR models and present promising results. For VAR, which requires 10-step generation (680 tokens), DD enables one-step generation (6.3$\times$ speed-up), with an acceptable increase in FID from 4.19 to 9.96. Similarly, for LlamaGen, DD reduces generation from 256 steps to 1, achieving an 217.8$\times$ speed-up with a comparable FID increase from 4.11 to 11.35. In both cases, baseline methods completely fail with FID scores $>$100. As the first work to demonstrate the possibility of one-step generation for image AR models, DD challenges the prevailing notion that AR models are inherently slow, and opens up new opportunities for efficient AR generation. The code and the pre-trained models will be released at https://github.com/imagination-research/distilled-decoding. The project website is at https://imagination-research.github.io/distilled-decoding.

1581. Accelerating Auto-regressive Text-to-Image Generation with Training-free Speculative Jacobi Decoding

链接：https://iclr.cc/virtual/2025/poster/29988 abstract： The current large auto-regressive models can generate high-quality, high-resolution images, but these models require hundreds or even thousands of steps of next-token prediction during inference, resulting in substantial time consumption. In existing studies, Jacobi decoding, an iterative parallel decoding algorithm, has been used to accelerate the auto-regressive generation and can be executed without training. However, the Jacobi decoding relies on a deterministic criterion to determine the convergence of iterations. Thus, it works for greedy decoding but is incompatible with sampling-based decoding which is crucial for visual quality and diversity in the current auto-regressive text-to-image generation. In this paper, we propose a training-free probabilistic parallel decoding algorithm, Speculative Jacobi Decoding (SJD), to accelerate auto-regressive text-to-image generation. By introducing a probabilistic convergence criterion, our SJD accelerates the inference of auto-regressive text-to-image generation while maintaining the randomness in sampling-based token decoding

and allowing the model to generate diverse images. Specifically, SJD facilitates the model to predict multiple tokens at each step and accepts tokens based on the probabilistic criterion, enabling the model to generate images with fewer steps than the conventional next-token-prediction paradigm. We also investigate the token initialization strategies that leverage the spatial locality of visual data to further improve the acceleration ratio under specific scenarios. We conduct experiments for our proposed SJD on multiple auto-regressive text-to-image generation models, showing the effectiveness of model acceleration without sacrificing the visual quality. The code of our work is available here: https://github.com/tyshiwo1/Accelerating-T2I-AR-with-SJD/.

1582. Linear Combination of Saved Checkpoints Makes Consistency and Diffusion Models Better

链接：https://iclr.cc/virtual/2025/poster/29672 abstract： Diffusion Models (DM) and Consistency Models (CM) are two types of popular generative models with good generation quality on various tasks. When training DM and CM, intermediate weight checkpoints are not fully utilized and only the last converged checkpoint is used. In this work, we find proper checkpoint merging can significantly improve the training convergence and final performance. Specifically, we propose LCSC, a simple but effective and efficient method to enhance the performance of DM and CM, by combining checkpoints along the training trajectory with coefficients deduced from evolutionary search. We demonstrate the value of LCSC through two use cases: (a) Reducing training cost. With LCSC, we only need to train DM/CM with fewer number of iterations and/or lower batch sizes to obtain comparable sample quality with the fully trained model. For example, LCSC achieves considerable training speedups for CM (23$\times$ on CIFAR-10 and 15$\times$ on ImageNet-64). (b) Enhancing pre-trained models. When full training is already done, LCSC can further improve the generation quality or efficiency of the final converged models. For example, LCSC achieves better FID using 1 number of function evaluation (NFE) than the base model with 2 NFE on consistency distillation, and decreases the NFE of DM from 15 to 9 while maintaining the generation quality. Applying LCSC to large text-to-image models, we also observe clearly enhanced generation quality.

1583. An Efficient Framework for Crediting Data Contributors of Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/30712 abstract： As diffusion models are deployed in real-world settings and their performance driven by training data, appraising the contribution of data contributors is crucial to creating incentives for sharing quality data and to implementing policies for data compensation. Depending on the use case, model performance corresponds to various global properties of the distribution learned by a diffusion model (e.g., overall aesthetic quality). Hence, here we address the problem of attributing global properties of diffusion models to data contributors. The Shapley value provides a principled approach to valuation by uniquely satisfying game-theoretic axioms of fairness. However, estimating Shapley values for diffusion models is computationally impractical because it requires retraining and rerunning inference on many subsets of data contributors. We introduce a method to efficiently retrain and rerun inference for Shapley value estimation, by leveraging model pruning and fine-tuning. We evaluate the utility of our method with three use cases: (i) image quality for a DDPM trained on a CIFAR dataset, (ii) demographic diversity for an LDM trained on CelebA-HQ, and (iii) aesthetic quality for a Stable Diffusion model LoRA-finetuned on Post-Impressionist artworks. Our results empirically demonstrate that our framework can identify important data contributors across global properties, outperforming existing attribution methods for diffusion models.

1584. TexTailor: Customized Text-aligned Texturing via Effective Resampling

链接：https://iclr.cc/virtual/2025/poster/31205 abstract： We present TexTailor, a novel method for generating consistent object textures from textual descriptions. Existing text-to-texture synthesis approaches utilize depth-aware diffusion models to progressively generate images and synthesize textures across predefined multiple viewpoints. However, these approaches lead to a gradual shift in texture properties across viewpoints due to (1) insufficient integration of previously synthesized textures at each viewpoint during the diffusion process and (2) the autoregressive nature of the texture synthesis process. Moreover, the predefined selection of camera positions, which does not account for the object's geometry, limits the effective use of texture information synthesized from different viewpoints, ultimately degrading overall texture consistency. In TexTailor, we address these issues by (1) applying a resampling scheme that repeatedly integrates information from previously synthesized textures within the diffusion process, and (2) fine-tuning a depth-aware diffusion model on these resampled textures. During this process, we observed that using only a few training images restricts the model's original ability to generate high-fidelity images aligned with the conditioning, and therefore propose an performance preservation loss to mitigate this issue. Additionally, we improve the synthesis of view-consistent textures by adaptively adjusting camera positions based on the object's geometry. Experiments on a subset of the Objaverse dataset and the ShapeNet car dataset demonstrate that TexTailor outperforms state-of-the-art methods in synthesizing view-consistent textures.

1585. ViDiT-Q: Efficient and Accurate Quantization of Diffusion Transformers for Image and Video Generation

链接：https://iclr.cc/virtual/2025/poster/30429 abstract： Diffusion transformers have demonstrated remarkable performance in visual generation tasks, such as generating realistic images or videos based on textual instructions. However, larger model sizes and multi-frame processing for video generation lead to increased computational and memory costs, posing challenges

for practical deployment on edge devices. Post-Training Quantization (PTQ) is an effective method for reducing memory costs and computational complexity.When quantizing diffusion transformers, we find that existing quantization methods face challenges when applied to text-to-image and video tasks. To address these challenges, we begin by systematically analyzing the source of quantization error and conclude with the unique challenges posed by DiT quantization. Accordingly, we design an improved quantization scheme: ViDiT-Q (Video \& Image Diffusion Transformer Quantization), tailored specifically for DiT models. We validate the effectiveness of ViDiT-Q across a variety of text-to-image and video models, achieving W8A8 and W4A8 with negligible degradation in visual quality and metrics. Additionally, we implement efficient GPU kernels to achieve practical 2-2.5x memory optimization and a 1.4-1.7x end-to-end latency speedup.

1586. A Large-scale Training Paradigm for Graph Generative Models

链接：https://iclr.cc/virtual/2025/poster/29077 abstract： Large Generative Models (LGMs) such as GPT, Stable Diffusion, Sora, and Suno are trained on a huge amount of texts, images, videos, and audio that are extremely diverse from numerous domains. This large-scale training paradigm on diverse well-curated data enhances the creativity and diversity of the generated content. However, all previous graph-generative models (e.g., GraphRNN, MDVAE, MoFlow, GDSS, and DiGress) have been trained only on one dataset each time, which cannot replicate the revolutionary success achieved by LGMs in other fields. To remedy this crucial gap, we propose a large-scale training paradigm that uses a large corpus of graphs (over 5000 graphs) from 13 domains, leading to the development of large graph generative models (LGGMs). We empirically demonstrate that the pre-trained LGGMs have superior zero-shot generative capability to existing graph generative models. Furthermore, our pre-trained LGGMs can be easily fine-tuned with graphs from target domains and demonstrate even better performance than those directly trained from scratch, behaving as a solid starting point for real-world customization. Inspired by Stable Diffusion, we further equip LGGMs with the Text-to-Graph generation capability, such as providing the description of the network name and domain (i.e., "The power-1138-bus graph represents a network of buses in a power distribution system.") and network statistics (i.e., "The graph has a low average degree, suitable for modeling social media interactions."). This Text-to-Graph capability integrates the extensive world knowledge in the underlying language model, offering users fine-grained control of the generated graphs. We release the code, the model checkpoint, and the datasets at https://github.com/KINDLab-Fly/LGGM.

1587. PooDLe□: Pooled and dense self-supervised learning from naturalistic videos

链接：https://iclr.cc/virtual/2025/poster/28997 abstract： Self-supervised learning has driven significant progress in learning from single-subject, iconic images.However, there are still unanswered questions about the use of minimally-curated, naturalistic video data, which contain dense scenes with many independent objects, imbalanced class distributions, and varying object sizes.In this paper, we propose PooDLe, a self-supervised learning method that combines an invariance-based objective on pooled representations with a dense SSL objective that enforces equivariance to optical flow warping.Our results show that a unified objective applied at multiple feature scales is essential for learning effective image representations from naturalistic videos.We validate our method with experiments on the BDD100K driving video dataset and the Walking Tours first-person video dataset, demonstrating its ability to capture spatial understanding from a dense objective and semantic understanding via a pooled representation objective.

1588. From GNNs to Trees: Multi-Granular Interpretability for Graph Neural Networks

链接：https://iclr.cc/virtual/2025/poster/30070 abstract： Interpretable Graph Neural Networks (GNNs) aim to reveal the underlying reasoning behind model predictions, attributing their decisions to specific subgraphs that are informative. However, existing subgraph-based interpretable methods suffer from an overemphasis on local structure, potentially overlooking long-range dependencies within the entire graphs. Although recent efforts that rely on graph coarsening have proven beneficial for global interpretability, they inevitably reduce the graphs to a fixed granularity. Such an inflexible way can only capture graph connectivity at a specific level, whereas real-world graph tasks often exhibit relationships at varying granularities (e.g., relevant interactions in proteins span from functional groups, to amino acids, and up to protein domains). In this paper, we introduce a novel Tree-like Interpretable Framework (TIF) for graph classification, where plain GNNs are transformed into hierarchical trees, with each level featuring coarsened graphs of different granularity as tree nodes. Specifically, TIF iteratively adopts a graph coarsening module to compress original graphs (i.e., root nodes of trees) into increasingly coarser ones (i.e., child nodes of trees), while preserving diversity among tree nodes within different branches through a dedicated graph perturbation module. Finally, we propose an adaptive routing module to identify the most informative root-to-leaf paths, providing not only the final prediction but also the multi-granular interpretability for the decision-making process. Extensive experiments on the graph classification benchmarks with both synthetic and real-world datasets demonstrate the superiority of TIF in interpretability, while also delivering a competitive prediction performance akin to the state-of-the-art counterparts.

1589. Stable Segment Anything Model

链接：https://iclr.cc/virtual/2025/poster/28332 abstract： The Segment Anything Model (SAM) achieves remarkable promptable segmentation given high-quality prompts which, however, often require good skills to specify. To make SAM robust to casual prompts, this paper presents the first comprehensive analysis on SAM’s segmentation stability across a diverse spectrum of prompt qualities, notably imprecise bounding boxes and insufficient points. Our key finding reveals that given such low-quality

prompts, SAM’s mask decoder tends to activate image features that are biased towards the background or confined to specific object parts. To mitigate this issue, our key idea consists of calibrating solely SAM’s mask attention by adjusting the sampling locations and amplitudes of image features, while the original SAM model architecture and weights remain unchanged. Consequently, our deformable sampling plugin (DSP) enables SAM to adaptively shift attention to the prompted target regions in a data-driven manner. During inference, dynamic routing plugin (DRP) is proposed that toggles SAM between the deformable and regular grid sampling modes, conditioned on the input prompt quality. Thus, our solution, termed Stable-SAM, offers several advantages: 1) improved SAM’s segmentation stability across a wide range of prompt qualities, while 2) retaining SAM’s powerful promptable segmentation efficiency and generality, with 3) minimal learnable parameters (0.08 M) and fast adaptation. Extensive experiments validate the effectiveness and advantages of our approach, underscoring Stable-SAM as a more robust solution for segmenting anything. Codes are at https://github.com/fanq15/Stable-SAM.

1590. Attention as a Hypernetwork

链接：https://iclr.cc/virtual/2025/poster/29433 abstract： Transformers can under some circumstances generalize to novel problem instances whose constituent parts might have been encountered during training, but whose compositions have not.What mechanisms underlie this ability for compositional generalization?By reformulating multi-head attention as a hypernetwork, we reveal that a composable, low-dimensional latent code specifies key-query specific operations.We find empirically that this latent code is predictive of the subtasks the network performs on unseen task compositions, revealing that latent codes acquired during training are reused to solve unseen problem instances.To further examine the hypothesis that the intrinsic hypernetwork of multi-head attention supports compositional generalization, we ablate whether making the hypernetwork-generated linear value network nonlinear strengthens compositionality.We find that this modification improves compositional generalization on abstract reasoning tasks.In particular, we introduce a symbolic version of the Raven's Progressive Matrices human intelligence test, which gives us precise control over the problem compositions encountered during training and evaluation.We demonstrate on this task how scaling model size and data enables compositional generalization in transformers and gives rise to a functionally structured latent space.

1591. Node Similarities under Random Projections: Limits and Pathological Cases

链接：https://iclr.cc/virtual/2025/poster/30323 abstract： Random Projections have been widely used to generate embeddings for various graph learning tasks due to their computational efficiency. The majority of applications have been justified through the Johnson-Lindenstrauss Lemma. In this paper, we take a step further and investigate how well dot product and cosine similarity are preserved by random projections when these are applied over the rows of the graph matrix. Our analysis provides new asymptotic and finite-sample results, identifies pathological cases, and tests them with numerical experiments. We specialize our fundamental results to a ranking application by computing the probability of random projections flipping the node ordering induced by their embeddings. We find that, depending on the degree distribution, the method produces especially unreliable embeddings for the dot product, regardless of whether the adjacency or the normalized transition matrix is used. With respect to the statistical noise introduced by random projections, we show that cosine similarity produces remarkably more precise approximations.

1592. OCEAN: Offline Chain-of-thought Evaluation and Alignment in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28173 abstract： Offline evaluation of LLMs is crucial in understanding their capacities, though current methods remain underexplored in existing research. In this work, we focus on the offline evaluation of the chain-of-thought capabilities and show how to optimize LLMs based on the proposed evaluation method. To enable offline feedback with rich knowledge and reasoning paths, we use knowledge graphs (KGs) (e.g., Wikidata5M) to provide feedback on the generated chain of thoughts. Due to the heterogeneity between LLM reasoning and KG structures, direct interaction and feedback from knowledge graphs on LLM behavior are challenging, as they require accurate entity linking and grounding of LLM-generated chains of thought in the KG. To address the above challenge, we propose an offline chain-of-thought evaluation framework, OCEAN, which models chain-of-thought reasoning in LLMs as a Markov Decision Process (MDP), and evaluate the policy’s alignment with KG preference modeling. To overcome the reasoning heterogeneity and grounding problems, we leverage on-policy KG exploration and reinforcement learning to model a KG policy that generates token-level likelihood distributions for LLM-generated chain-of-thought reasoning paths, simulating KG reasoning preference. Then we incorporate the knowledge-graph feedback on the validity and alignment of the generated reasoning paths into inverse propensity scores and propose KG-IPS estimator. Theoretically, we prove the unbiasedness of the proposed KG-IPS estimator and provide a lower bound on its variance. With the off-policy evaluated value function, we can directly enable off-policy optimization to further enhance chain-of-thought alignment. Our empirical study shows that OCEAN can be efficiently optimized for generating chain-of-thought reasoning paths with higher estimated values without affecting LLMs’ general abilities in downstream tasks or their internal knowledge.

1593. Permute-and-Flip: An optimally stable and watermarkable decoder for LLMs

链接：https://iclr.cc/virtual/2025/poster/29239 abstract： In this paper, we propose a new decoding method called Permute-

and-Flip (PF) decoder. It enjoys stability properties similar to the standard sampling decoder, but is provably up to 2x better in its quality-stability tradeoff than sampling and never worse than any other decoder. We also design a cryptographic watermarking scheme analogous to Aaronson (2023)'s Gumbel watermark, but naturally tailored for PF decoder. The watermarking scheme does not change the distribution to sample, while allowing arbitrarily low false positive rate and high recall whenever the generated text has high entropy. Our experiments show that the PF decoder (and its watermarked counterpart) significantly outperform(s) naive sampling (and its Gumbel watermarked counterpart) in terms of perplexity, while retaining the same stability (and detectability), hence making it a promising new approach for LLM decoding. The code is available at https://github.com/XuandongZhao/pf-decoding

1594. Breaking Class Barriers: Efficient Dataset Distillation via Inter-Class Feature Compensator

链接：https://iclr.cc/virtual/2025/poster/29331 abstract： Dataset distillation has emerged as a technique aiming to condense informative features from large, natural datasets into a compact and synthetic form. While recent advancements have refined this technique, its performance is bottlenecked by the prevailing class-specific synthesis paradigm. Under this paradigm, synthetic data is optimized exclusively for a pre-assigned one-hot label, creating an implicit class barrier in feature condensation. This leads to inefficient utilization of the distillation budget and oversight of inter-class feature distributions, which ultimately limits the effectiveness and efficiency, as demonstrated in our analysis.To overcome these constraints, this paper presents the Inter-class Feature Compensator (INFER), an innovative distillation approach that transcends the class-specific data-label framework widely utilized in current dataset distillation methods. Specifically, INFER leverages a Universal Feature Compensator (UFC) to enhance feature integration across classes, enabling the generation of multiple additional synthetic instances from a single UFC input. This significantly improves the efficiency of the distillation budget.Moreover, INFER enriches inter-class interactions during the distillation, thereby enhancing the effectiveness and generalizability of the distilled data. By allowing for the linear interpolation of labels similar to those in the original dataset, INFER meticulously optimizes the synthetic data and dramatically reduces the size of soft labels in the synthetic dataset to almost zero, establishing a new benchmark for efficiency and effectiveness in dataset distillation. In practice, INFER demonstrates state-of-the-art performance across benchmark datasets. For instance, in the $\texttt{ipc} = 50$ setting on ImageNet-1k with the same compression level, it outperforms SRe2L by 34.5\% using ResNet18. Codes are available at https://github.com/zhangxin-xd/UFC.

1595. STORM: Spatio-TempOral Reconstruction Model For Large-Scale Outdoor Scenes

链接：https://iclr.cc/virtual/2025/poster/29959 abstract： We present STORM, a spatio-temporal reconstruction model designed for reconstructing dynamic outdoor scenes from sparse observations. Existing dynamic reconstruction methods often rely on per-scene optimization, dense observations across space and time, and strong motion supervision, resulting in lengthy optimization times, limited generalization to novel views or scenes, and degenerated quality caused by noisy pseudo-labels for dynamics. To address these challenges, STORM leverages a data-driven Transformer architecture that directly infers dynamic 3D scene representations—parameterized by 3D Gaussians and their velocities—in a single forward pass. Our key design is to aggregate 3D Gaussians from all frames using self-supervised scene flows, transforming them to the target timestep to enable complete (i.e., "amodal") reconstructions from arbitrary viewpoints at any moment in time. As an emergent property, STORM automatically captures dynamic instances and generates high-quality masks using only reconstruction losses. Extensive experiments on public datasets show that STORM achieves precise dynamic scene reconstruction, surpassing state-of-the-art per-scene optimization methods (+4.3 to 6.6 PSNR) and existing feed-forward approaches (+2.1 to 4.7 PSNR) in dynamic regions. STORM reconstructs large-scale outdoor scenes in 200ms, supports real-time rendering, and outperforms competitors in scene flow estimation, improving 3D EPE by 0.422m and Acc5 by 28.02%. Beyond reconstruction, we showcase four additional applications of our model, illustrating the potential of self-supervised learning for broader dynamic scene understanding. For more details, please visit our project at https://jiawei-yang.github.io/STORM/.

1596. UniGEM: A Unified Approach to Generation and Property Prediction for Molecules

链接：https://iclr.cc/virtual/2025/poster/29987 abstract： Molecular generation and molecular property prediction are both crucial for drug discovery, but they are often developed independently. Inspired by recent studies, which demonstrate that diffusion model, a prominent generative approach, can learn meaningful data representations that enhance predictive tasks, we explore the potential for developing a unified generative model in the molecular domain that effectively addresses both molecular generation and property prediction tasks. However, the integration of these tasks is challenging due to inherent inconsistencies, making simple multi-task learning ineffective. To address this, we propose UniGEM, the first unified model to successfully integrate molecular generation and property prediction, delivering superior performance in both tasks. Our key innovation lies in a novel two-phase generative process, where predictive tasks are activated in the later stages, after the molecular scaffold is formed. We further enhance task balance through innovative training strategies. Rigorous theoretical analysis and comprehensive experiments demonstrate our significant improvements in both tasks. The principles behind UniGEM hold promise for broader applications, including natural language processing and computer vision.

1597. Harnessing Webpage UIs for Text-Rich Visual Understanding

链接：https://iclr.cc/virtual/2025/poster/30176 abstract： Text-rich visual understanding—the ability to interpret both textual content and visual elements within a scene—is crucial for multimodal large language models (MLLMs) to effectively interact with structured environments. We propose leveraging webpage UIs as a naturally structured and diverse data source to enhance MLLMs' capabilities in this area. Existing approaches, such as rule-based extraction, multimodal model captioning, and rigid HTML parsing, are hindered by issues like noise, hallucinations, and limited generalization. To overcome these challenges, we introduce MultiUI, a dataset of 7.3 million samples spanning various UI types and tasks, structured using enhanced accessibility trees and task taxonomies. By scaling multimodal instructions from web UIs through LLMs, our dataset enhances generalization beyond web domains, significantly improving performance in document understanding, GUI comprehension, grounding, and advanced agent tasks. This demonstrates the potential of structured web data to elevate MLLMs' proficiency in processing text-rich visual environments and generalizing across domains.

1598. Spider 2.0: Evaluating Language Models on Real-World Enterprise Text-to-SQL Workflows

链接：https://iclr.cc/virtual/2025/poster/29294 abstract： Real-world enterprise text-to-SQL workflows often involve complex cloud or local data across various database systems, multiple SQL queries in various dialects, and diverse operations from data transformation to analytics.We introduce Spider 2.0, an evaluation framework comprising 632 real-world text-to-SQL workflow problems derived from enterprise-level database use cases. The databases in Spider 2.0 are sourced from real data applications, often containing over 1,000 columns and stored in local or cloud database systems such as BigQuery and Snowflake.We show that solving problems in Spider 2.0 frequently requires understanding and searching through database metadata, dialect documentation, and even project-level codebases. This challenge calls for models to interact with complex SQL workflow environments, process extremely long contexts, perform intricate reasoning, and generate multiple SQL queries with diverse operations, often exceeding 100 lines, which goes far beyond traditional text-to-SQL challenges.Our evaluations indicate that based on o1-preview, our code agent framework successfully solves only 21.3\% of the tasks, compared with 91.2\% on Spider 1.0 and 73.0\% on BIRD.Our results on Spider 2.0 show that while language models have demonstrated remarkable performance in code generation --- especially in prior text-to-SQL benchmarks --- they require significant improvement in order to achieve adequate performance for real-world enterprise usage.Progress on Spider 2.0 represents crucial steps towards developing intelligent, autonomous, code agents for real-world enterprise settings.Our code, baseline models, and data are available at spider2-sql.github.io .

1599. Prompt as Knowledge Bank: Boost Vision-language model via Structural Representation for zero-shot medical detection

链接：https://iclr.cc/virtual/2025/poster/28543 abstract： Zero-shot medical detection can further improve detection performance without relying on annotated medical images even upon the fine-tuned model, showing great clinical value. Recent studies leverage grounded vision-language models (GLIP) to achieve this by using detailed disease descriptions as prompts for the target disease name during the inference phase. However, these methods typically treat prompts as equivalent context to the target name, making it difficult to assign specific disease knowledge based on visual information, leading to a coarse alignment between images and target descriptions. In this paper, we propose StructuralGLIP, which introduces an auxiliary branch to encode prompts into a latent knowledge bank layer-by-layer, enabling more context-aware and fine-grained alignment. Specifically, in each layer, we select highly similar features from both the image representation and the knowledge bank, forming structural representations that capture nuanced relationships between image patches and target descriptions. These features are then fused across modalities to further enhance detection performance.Extensive experiments demonstrate that StructuralGLIP achieves a +4.1\% AP improvement over prior state-of-the-art methods across seven zero-shot medical detection benchmarks, and consistently improves fine-tuned models by +3.2\% AP on endoscopy image datasets.

1600. How do we interpret the outputs of a neural network trained on classification?

链接：https://iclr.cc/virtual/2025/poster/31365 abstract： Deep neural networks are widely used for classification tasks, but the interpretation of their output activations is often unclear. This post explains how these outputs can be understood as approximations of the Bayesian posterior probability. We showed that, in theory, the loss function for classification tasks -- derived by maximum likelihood -- is minimized by the Bayesian posterior. We conducted empirical studies training neural networks to classify synthetic data from a known generative model. In a simple classification task, the network closely approximates the theoretically derived posterior. However, simple changes in the task can make accurate approximation much more difficult. The model's ability to approximate the posterior depends on multiple factors, such as the complexity of the posterior and whether there is sufficient data for learning.