2601. Closed-Form Merging of Parameter-Efficient Modules for Federated Continual Learning

链接：https://iclr.cc/virtual/2025/poster/29651 abstract： Model merging has emerged as a crucial technique in Deep Learning, enabling the integration of multiple models into a unified system while preserving performance and scalability. In this respect, the compositional properties of low-rank adaptation techniques (e.g., LoRA) have proven beneficial, as simple averaging LoRA modules yields a single model that mostly integrates the capabilities of all individual modules. Building on LoRA, we take a step further by imposing that the merged model matches the responses of all learned modules. Solving this ob-jective in closed form yields an indeterminate system with A and B as unknown variables, indicating the existence of infinitely many closed-form solutions. To address this challenge, we introduce LoRM, an alternating optimization strategy that trains one LoRA matrix at a time. This allows solving for each unknown variable individually, thus finding a unique solution. We apply our proposed methodology to Federated Class-Incremental Learning (FCIL), ensuring alignment of model responses both between clients and across tasks. Our method demonstrates state-of-the-art performance across a range of FCIL scenarios. The code to reproduce our experiments is available at github.com/aimagelab/fed-mammoth.

2602. Noise Separation guided Candidate Label Reconstruction for Noisy Partial Label Learning

链接：https://iclr.cc/virtual/2025/poster/29547 abstract： Partial label learning is a weakly supervised learning problem in which an instance is annotated with a set of candidate labels, among which only one is the correct label. However, in practice the correct label is not always in the candidate label set, leading to the noisy partial label learning (NPLL) problem. In this paper, we theoretically prove that the generalization error of the classifier constructed under NPLL paradigm is bounded by the noise rate and the average length of the candidate label set. Motivated by the theoretical guide, we propose a novel NPLL framework that can separate the noisy samples from the normal samples to reduce the noise rate and reconstruct the shorter candidate label sets for both of them. Extensive experiments on multiple benchmark datasets confirm the efficacy of the proposed method in addressing NPLL. For example, on CIFAR100 dataset with severe noise, our method improves the classification accuracy of the state-of-the-art one by 11.57%. The code is available at: https://github.com/pruirui/PLRC.

2603. ILLUSION: Unveiling Truth with a Comprehensive Multi-Modal, Multi-Lingual Deepfake Dataset

链接：https://iclr.cc/virtual/2025/poster/28235 abstract： The proliferation of deepfakes and AI-generated content has led to a surge in media forgeries and misinformation, necessitating robust detection systems. However, current datasets lack diversity across modalities, languages, and real-world scenarios. To address this gap, we present ILLUSION (Integration of Life-Like Unique Synthetic Identities and Objects from Neural Networks), a large-scale, multi-modaldeepfake dataset comprising 1.3 million samples spanning audio-visual forgeries, 26 languages, challenging noisy environments, and various manipulation protocols. Generated using 28 state-of-the-art generative techniques, ILLUSION includesfaceswaps, audio spoofing, synchronized audio-video manipulations, and synthetic media while ensuring a balanced representation of gender and skin tone for unbiased evaluation. Using Jaccard Index and UpSet plot analysis, we demonstrate ILLUSION's distinctiveness and minimal overlap with existing datasets, emphasizing its novel generative coverage. We benchmarked image, audio, video, and multi-modal detection models, revealing key challenges such as performance degradation in multilingual and multi-modal contexts, vulnerability to real-world distortions, and limited generalization to zero-day attacks. By bridging synthetic and real-world complexities, ILLUSION provides a challenging yet essential platform for advancing deepfake detection research. The dataset is publicly available at https://www.iab-rubric.org/illusion-database.

2604. MLLM can see? Dynamic Correction Decoding for Hallucination Mitigation

链接：https://iclr.cc/virtual/2025/poster/30978 abstract： Multimodal Large Language Models (MLLMs) frequently exhibit hallucination phenomena, but the underlying reasons remain poorly understood. In this paper, we present an empirical analysis and find that, although MLLMs incorrectly generate the objects in the final output, they are actually able to recognize visual objects in the preceding layers. We speculate that this may be due to the strong knowledge priors of the language model suppressing the visual information, leading to hallucinations. Motivated by this, we propose a novel dynamic correction decoding method for MLLMs DeCo, which adaptively selects the appropriate preceding layers and proportionally integrates knowledge into the final layer to adjust the output logits. Note that DeCo is model agnostic and can be seamlessly incorporated with various classic decoding strategies and applied to different MLLMs. We evaluate DeCo on widely-used benchmarks, demonstrating that it can reduce hallucination rates by a large margin compared to baselines, highlighting its potential to mitigate hallucinations. Code is available at https://github.com/zjunlp/DeCo.

2605. Boosting Multiple Views for pretrained-based Continual Learning

链接：https://iclr.cc/virtual/2025/poster/30627 abstract： Recent research has shown that Random Projection (RP) can effectively improve the performance of pre-trained models in Continual learning (CL). The authors hypothesized that using RP to

map features onto a higher-dimensional space can make them more linearly separable. In this work, we theoretically analyze the role of RP and present its benefits for improving the model's generalization abilityin each task and facilitating CL overall. Additionally, we take this result to the next level by proposing a Multi-View Random Projection scheme for a stronger ensemble classifier. In particular, we train a set of linear experts, among which diversity is encouraged based on the principle of AdaBoost, which was initially very challenging to apply to CL. Moreover, we employ a task-based adaptive backbonewith distinct prompts dedicated to each task for better representation learning. To properly select these task-specific components and mitigate potential feature shifts caused by misprediction, we introduce a simple yet effective technique called the self-improvement process. Experimentally, our method consistently outperforms state-of-the-art baselines across a wide range of datasets.

2606. Neural ODE Transformers: Analyzing Internal Dynamics and Adaptive Fine-tuning

链接：https://iclr.cc/virtual/2025/poster/29292 abstract： Recent advancements in large language models (LLMs) based on transformer architectures have sparked significant interest in understanding their inner workings. In this paper, we introduce a novel approach to modeling transformer architectures using highly flexible non-autonomous neural ordinary differential equations (ODEs). Our proposed model parameterizes all weights of attention and feed-forward blocks through neural networks, expressing these weights as functions of a continuous layer index. Through spectral analysis of the model's dynamics, we uncover an increase in eigenvalue magnitude that challenges the weight-sharing assumption prevalent in existing theoretical studies. We also leverage the Lyapunov exponent to examine token-level sensitivity, enhancing model interpretability. Our neural ODE transformer demonstrates performance comparable to or better than vanilla transformers across various configurations and datasets, while offering flexible fine-tuning capabilities that can adapt to different architectural constraints.

2607. PaLD: Detection of Text Partially Written by Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28190 abstract： Advances in large language models (LLM) have produced text that appears increasingly human-like and difficult to detect with the human eye. In order to mitigate the impact of misusing LLM-generated texts, e.g., copyright infringement, fair student assessment, fraud, and other societally harmful LLM usage, a line of work on detecting human and LLM-written text has been explored. While recent work has focused on classifying entire text samples (e.g., paragraphs) as human or LLM-written, this paper investigates a more realistic setting of mixed-text, where the text's individual segments (e.g., sentences) could each be written by either a human or an LLM. A text encountered in practical usage cannot generally be assumed to be fully human or fully LLM-written; simply predicting whether it is human or LLM-written is insufficient as it does not provide the user with full context on its origins, such as the amount of LLM-written text, or locating the LLM-written parts. Therefore, we study two relevant problems in the mixed-text setting: (i) estimating the percentage of a text that was LLM-written, and (ii) determining which segments were LLM-written. To this end, we propose Partial-LLM Detector (PaLD), a black-box method that leverages the scores of text classifiers. Experimentally, we demonstrate the effectiveness of PaLD compared to baseline methods that build on existing LLM text detectors.

2608. Benchmarking Agentic Workflow Generation

链接：https://iclr.cc/virtual/2025/poster/27882 abstract： Large Language Models (LLMs), with their exceptional ability to handle a wide range of tasks, have driven significant advancements in tackling reasoning and planning tasks, wherein decomposing complex problems into executable workflows is a crucial step in this process. Existing workflow evaluation frameworks either focus solely on holistic performance or suffer from limitations such as restricted scenario coverage, simplistic workflow structures, and lax evaluation standards. To this end, we introduce WorfBench, a unified workflow generation benchmark with multi-faceted scenarios and intricate graph workflow structures. Additionally, we present WorfEval, a systemic evaluation protocol utilizing subsequence and subgraph matching algorithms to accurately quantify the LLM agent's workflow generation capabilities. Through comprehensive evaluations across different types of LLMs, we discover distinct gaps between the sequence planning capabilities and graph planning capabilities of LLM agents, with even GPT-4 exhibiting a gap of around 15%. We also train two open-source models and evaluate their generalization abilities on held-out tasks. Furthermore, we observe that the generated workflows can enhance downstream tasks, enabling them to achieve superior performance with less time during inference. Code and dataset are available at https://github.com/zjunlp/WorfBench.

2609. SAVA: Scalable Learning-Agnostic Data Valuation

链接：https://iclr.cc/virtual/2025/poster/31253 abstract： Selecting data for training machine learning models is crucial since large, web-scraped, real datasets contain noisy artifacts that affect the quality and relevance of individual data points. These noisy artifacts will impact model performance. We formulate this problem as a data valuation task, assigning a value to data points in the training set according to how similar or dissimilar they are to a clean and curated validation set. Recently, LAVA (Just et al., 2023) demonstrated the use of optimal transport (OT) between a large noisy training dataset and a clean validation set, to value training data efficiently, without the dependency on model performance. However, the LAVA algorithm requires the entire dataset as an input, this limits its application to larger datasets. Inspired by the scalability of stochastic (gradient) approaches which carry out computations on batches of data points instead of the entire dataset, we analogously propose SAVA, a scalable variant of LAVA with its computation on batches of data points. Intuitively, SAVA follows the same scheme as LAVA which leverages the hierarchically defined OT for data valuation. However, while LAVA processes the whole dataset, SAVA divides the dataset into batches of data points, and carries out the OT problem computation on those batches. Moreover, our theoretical derivations on the trade-off of using entropic regularization for OT problems include refinements of prior work. We

perform extensive experiments, to demonstrate that SAVA can scale to large datasets with millions of data points and does not trade off data valuation performance. Our Github repository is available at \url{https://github.com/skezle/sava}.

2610. ADMM for Structured Fractional Minimization

链接：https://iclr.cc/virtual/2025/poster/30451 abstract： This paper considers a class of structured fractional minimization problems. The numerator consists of a differentiable function, a simple nonconvex nonsmooth function, a concave nonsmooth function, and a convex nonsmooth function composed with a linear operator. The denominator is a continuous function that is either weakly convex or has a weakly convex square root. These problems are prevalent in various important applications in machine learning and data science. Existing methods, primarily based on subgradient methods and smoothing proximal gradient methods, often suffer from slow convergence and numerical stability issues. In this paper, we introduce {\sf FADMM}, the first Alternating Direction Method of Multipliers tailored for this class of problems. {\sf FADMM} decouples the original problem into linearized proximal subproblems, featuring two variants: one using Dinkelbach's parametric method ({\sf FADMM-D}) and the other using the quadratic transform method ({\sf FADMM-Q}). By introducing a novel Lyapunov function, we establish that {\sf FADMM} converges to ϵ-approximate critical points of the problem within an oracle complexity of $\mathcal{O}(1/\epsilon^{3})$. Extensive experiments on synthetic and real-world datasets, including sparse Fisher discriminant analysis, robust Sharpe ratio minimization, and robust sparse recovery, demonstrate the effectiveness of our approach.

2611. Beyond Interpretability: The Gains of Feature Monosemanticity on Model Robustness

链接：https://iclr.cc/virtual/2025/poster/28835 abstract： Deep learning models often suffer from a lack of interpretability due to \emph{polysemanticity}, where individual neurons are activated by multiple unrelated semantics, resulting in unclear attributions of model behavior. Recent advances in \emph{monosemanticity}, where neurons correspond to consistent and distinct semantics, have significantly improved interpretability but are commonly believed to compromise accuracy. In this work, we challenge the prevailing belief of the accuracy-interpretability tradeoff, showing that monosemantic features not only enhance interpretability but also bring concrete gains in model performance of {\color{black} robustness-related tasks}. Across multiple robust learning scenarios—including input and label noise, few-shot learning, and out-of-domain generalization—our results show that models leveraging monosemantic features significantly outperform those relying on polysemantic features. Furthermore, we provide empirical and theoretical understandings on the robustness gains of feature monosemanticity. Our preliminary analysis suggests that monosemanticity, by promoting better separation of feature representations, leads to more robust decision boundaries {\color{black} under noise}. This diverse evidence highlights the \textbf{generality} of monosemanticity in improving model robustness. As a first step in this new direction, we embark on exploring the learning benefits of monosemanticity beyond interpretability, supporting the long-standing hypothesis of linking interpretability and robustness. Code is available at \url{https://github.com/PKU-ML/Monosemanticity-Robustness}.

2612. Projection Head is Secretly an Information Bottleneck

链接：https://iclr.cc/virtual/2025/poster/30021 abstract： Recently, contrastive learning has risen to be a promising paradigm for extracting meaningful data representations. Among various special designs, adding a projection head on top of the encoder during training and removing it for downstream tasks has proven to significantly enhance the performance of contrastive learning. However, despite its empirical success, the underlying mechanism of the projection head remains under-explored. In this paper, we develop an in-depth theoretical understanding of the projection head from the information-theoretic perspective. By establishing the theoretical guarantees on the downstream performance of the features before the projector, we reveal that an effective projector should act as an information bottleneck, filtering out the information irrelevant to the contrastive objective. Based on theoretical insights, we introduce modifications to projectors with training and structural regularizations. Empirically, our methods exhibit consistent improvement in the downstream performance across various real-world datasets, including CIFAR-10, CIFAR-100, and ImageNet-100. We believe our theoretical understanding on the role of the projection head will inspire more principled and advanced designs in this field. Code is available at \url{https://github.com/PKU-ML/Projector_Theory}.

2613. Boltzmann Semantic Score: A Semantic Metric for Evaluating Large Vision Models Using Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30657 abstract： Do Large Vision Models (LVMs) extract medically and semantically relevant features similar to those identified by human experts? Currently, only biased, qualitative approaches with limited, small-scale expert evaluations are available to answer this question. In this study, we propose the Boltzmann Semantic Score (BSS), a novel method inspired by state space modeling, to evaluate the encoding space of LVMs from medical images using the encoding space of Large Language Models (LLMs) from medical reports. Through extensive experimentation on 32 datasets from The Cancer Genome Atlas collection using five state-of-the-art LLMs, we first establish a baseline of LLMs' performance in digital pathology and show that LLMs' encoding can be linked to patient outcomes. Then, we compared seven LVMs with BSS and showed that LVMs suffer from poor semantic capability when compared with encoded expert knowledge from pathology reports.We also found statistically significant correlations between BSS (as a measure of structural similarity) and performance in two downstream tasks: information retrieval and survival prediction tasks. Our study also investigates the consensus among LLMs in evaluating LVMs using BSS, indicating that LLMs generally reach substantial consensus in rating LVMs, with some

variation dependant on the cancer type. We believe the BSS metric proposed here holds significant potential for application in other domains with similar contexts. Data and code can be found in \footnotesize \url{ https://github.com/AIMLab-UBC/Boltzmann}

2614. Unsupervised Zero-Shot Reinforcement Learning via Dual-Value Forward-Backward Representation

链接：https://iclr.cc/virtual/2025/poster/31255 abstract： Online unsupervised reinforcement learning (URL) can discover diverse skills via reward-free pre-training and exhibits impressive downstream task adaptation abilities through further fine-tuning.However, online URL methods face challenges in achieving zero-shot generalization, i.e., directly applying pre-trained policies to downstream tasks without additional planning or learning.In this paper, we propose a novel Dual-Value Forward-Backward representation (DVFB) framework with a contrastive entropy intrinsic reward to achieve both zero-shot generalization and fine-tuning adaptation in online URL.On the one hand, we demonstrate that poor exploration in forward-backward representations can lead to limited data diversity in online URL, impairing successor measures, and ultimately constraining generalization ability.To address this issue, the DVFB framework learns successor measures through a skill value function while promoting data diversity through an exploration value function, thus enabling zero-shot generalization.On the other hand, and somewhat surprisingly, by employing a straightforward dual-value fine-tuning scheme combined with a reward mapping technique, the pre-trained policy further enhances its performance through fine-tuning on downstream tasks, building on its zero-shot performance.Through extensive multi-task generalization experiments, DVFB demonstrates both superior zero-shot generalization (outperforming on all 12 tasks) and fine-tuning adaptation (leading on 10 out of 12 tasks) abilities, surpassing state-of-the-art URL methods.

2615. LoCoDL: Communication-Efficient Distributed Learning with Local Training and Compression

链接：https://iclr.cc/virtual/2025/poster/29728 abstract： In Distributed optimization and Learning, and even more in the modern framework of federated learning, communication, which is slow and costly, is critical. We introduce LoCoDL, a communication-efficient algorithm that leverages the two popular and effective techniques of Local training, which reduces the communication frequency, and Compression, in which short bitstreams are sent instead of full-dimensional vectors of floats. LoCoDL works with a large class of unbiased compressors that includes widely-used sparsification and quantization methods. LoCoDL provably benefits from local training and compression and enjoys a doubly-accelerated communication complexity, with respect to the condition number of the functions and the model dimension, in the general heterogeneous regime with strongly convex functions. This is confirmed in practice, with LoCoDL outperforming existing algorithms.

2616. IDA-VLM: Towards Movie Understanding via ID-Aware Large Vision-Language Model

链接：https://iclr.cc/virtual/2025/poster/32091 abstract： The rapid advancement of Large Vision-Language models (LVLMs) has demonstrated a spectrum of emergent capabilities. Nevertheless, current models only focus on the visual content of a single scenario, while their ability to associate instances across different scenes has not yet been explored, which is essential for understanding complex visual content, such as movies with multiple characters and intricate plots. Towards movie understanding, a critical initial step for LVLMs is to unleash the potential of character identities memory and recognition across multiple visual scenarios. To achieve the goal, we propose visual instruction tuning with ID reference and develop an ID-Aware Large Vision-Language Model, IDA-VLM. Furthermore, our research introduces a novel benchmark MM-ID, to examine LVLMs on instance IDs memory and recognition across four dimensions: matching, location, question-answering, and captioning. Our findings highlight the limitations of existing LVLMs in recognizing and associating instance identities with ID reference. This paper paves the way for future artificial intelligence systems to possess multi-identity visual inputs, thereby facilitating the comprehension of complex visual narratives like movies.

2617. XLand-100B: A Large-Scale Multi-Task Dataset for In-Context Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28316 abstract： Following the success of the in-context learning paradigm in large-scale language and computer vision models, the recently emerging field of in-context reinforcement learning is experiencing a rapid growth. However, its development has been held back by the lack of challenging benchmarks, as all the experiments have been carried out in simple environments and on small-scale datasets. We present **XLand-100B**, a large-scale dataset for in-context reinforcement learning based on the XLand-MiniGrid environment, as a first step to alleviate this problem. It contains complete learning histories for nearly $30,000$ different tasks, covering 100B transitions and 2.5B episodes. It took $50,000$ GPU hours to collect the dataset, which is beyond the reach of most academic labs. Along with the dataset, we provide the utilities to reproduce or expand it even further. We also benchmark common in-context RL baselines and show that they struggle to generalize to novel and diverse tasks. With this substantial effort, we aim to democratize research in the rapidly growing field of in-context reinforcement learning and provide a solid foundation for further scaling.

2618. MindSimulator: Exploring Brain Concept Localization via Synthetic fMRI

链接：https://iclr.cc/virtual/2025/poster/27898 abstract： Concept-selective regions within the human cerebral cortex exhibit significant activation in response to specific visual stimuli associated with particular concepts. Precisely localizing these regions stands as a crucial long-term goal in neuroscience to grasp essential brain functions and mechanisms. Conventional experiment-driven approaches hinge on manually constructed visual stimulus collections and corresponding brain activity recordings, constraining the support and coverage of concept localization. Additionally, these stimuli often consist of concept objects in unnatural contexts and are potentially biased by subjective preferences, thus prompting concerns about the validity and generalizability of the identified regions. To address these limitations, we propose a data-driven exploration approach. By synthesizing extensive brain activity recordings, we statistically localize various concept-selective regions. Our proposed MindSimulator leverages advanced generative technologies to learn the probability distribution of brain activity conditioned on concept-oriented visual stimuli. This enables the creation of simulated brain recordings that reflect real neural response patterns. Using the synthetic recordings, we successfully localize several well-studied concept-selective regions and validate them against empirical findings, achieving promising prediction accuracy. The feasibility opens avenues for exploring novel concept-selective regions and provides prior hypotheses for future neuroscience research.

2619. SELF-EVOLVED REWARD LEARNING FOR LLMS

链接：https://iclr.cc/virtual/2025/poster/29197 abstract： Reinforcement Learning from Human Feedback (RLHF) is a crucial technique for aligning language models with human preferences and is a key factor in the success of modern conversational models like GPT-4, ChatGPT, and Llama 2. A significant challenge in employing RLHF lies in training a reliable RM, which relies on high-quality labels. Typically, these labels are provided by human experts or a stronger AI, both of which can be costly and introduce bias that may affect the language model's responses. As models improve, human input may become less effective in enhancing their performance. This paper explores the potential of using the RM itself to generate additional training data for a more robust RM. Our experiments demonstrate that reinforcement learning from self-feedback outperforms baseline approaches.We conducted extensive experiments with our approach on multiple datasets, such as HH-RLHF and UltraFeedback, and models including Mistral and Llama 3, comparing it against various baselines. Our results indicate that, even with a limited amount of human-labeled data, learning from self-feedback can robustly enhance the performance of the RM, thereby improving the capabilities of large language models.

2620. RuAG: Learned-rule-augmented Generation for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30545 abstract： In-context learning (ICL) and Retrieval-Augmented Generation (RAG) have gained attention for their ability to enhance LLMs' reasoning by incorporating external knowledge but suffer from limited contextual window size, leading to insufficient information injection. To this end, we propose a novel framework to automatically distill large volumes of offline data into interpretable first-order logic rules, which are injected into LLMs to boost their reasoning capabilities. Our method begins by formulating the search process relying on LLMs' commonsense, where LLMs automatically define head and body predicates. Then, we apply Monte Carlo Tree Search (MCTS) to address the combinational searching space and efficiently discover logic rules from data. The resulting logic rules are translated into natural language, allowing targeted knowledge injection and seamless integration into LLM prompts for LLM's downstream task reasoning. We evaluate our framework on public and private industrial tasks, including Natural Language Processing (NLP), time-series, decision-making, and industrial tasks, demonstrating its effectiveness in enhancing LLM's capability over diverse tasks.

2621. On the Adversarial Vulnerability of Label-Free Test-Time Adaptation

链接：https://iclr.cc/virtual/2025/poster/29905 abstract： Despite the success of Test-time adaptation (TTA), recent work has shown that adding relatively small adversarial perturbations to a limited number of samples leads to significant performance degradation. Therefore, it is crucial to rigorously evaluate existing TTA algorithms against relevant threats and implement appropriate security countermeasures. Importantly, existing threat models assume test-time samples will be labeled, which is impractical in real-world scenarios. To address this gap, we propose a new attack algorithm that does not rely onaccess to labeled test samples, thus providing a concrete way to assess the security vulnerabilities of TTA algorithms. Our attack design is grounded in theoretical foundations and can generate strong attacks against different state of the art TTA methods. In addition, we show that existing defense mechanisms are almost ineffective, which emphasizes the need for further research on TTA security. Through extensive experiments on CIFAR10-C, CIFAR100-C, and ImageNet-C, we demonstrate that our proposed approach closely matches the performance of state-of-the-art attack benchmarks, even without access to labeled samples. In certain cases, our approach generates stronger attacks, e.g., more than 4% higher error rate on CIFAR10-C.

2622. IMDPrompter: Adapting SAM to Image Manipulation Detection by Cross-View Automated Prompt Learning

链接：https://iclr.cc/virtual/2025/poster/27891 abstract： Using extensive training data from SA-1B, the Segment Anything Model (SAM) has demonstrated exceptional generalization and zero-shot capabilities, attracting widespread attention in areas

such as medical image segmentation and remote sensing image segmentation. However, its performance in the field of image manipulation detection remains largely unexplored and unconfirmed. There are two main challenges in applying SAM to image manipulation detection: a) reliance on manual prompts, and b) the difficulty of single-view information in supporting cross-dataset generalization. To address these challenges, we develops a cross-view prompt learning paradigm called IMDPrompter based on SAM. Benefiting from the design of automated prompts, IMDPrompter no longer relies on manual guidance, enabling automated detection and localization. Additionally, we propose components such as Cross-view Feature Perception, Optimal Prompt Selection, and Cross-View Prompt Consistency, which facilitate cross-view perceptual learning and guide SAM to generate accurate masks. Extensive experimental results from five datasets (CASIA, Columbia, Coverage, IMD2020, and NIST16) validate the effectiveness of our proposed method.

2623. Autocorrelation Matters: Understanding the Role of Initialization Schemes for State Space Models

链接：https://iclr.cc/virtual/2025/poster/28120 abstract： Current methods for initializing state space model (SSM) parameters primarily rely on the HiPPO framework \citep{gu2023how}, which is based on online function approximation with the SSM kernel basis. However, the HiPPO framework does not explicitly account for the effects of the temporal structures of input sequences on the optimization of SSMs.In this paper, we take a further step to investigate the roles of SSM initialization schemes by considering the autocorrelation of input sequences. Specifically, we: (1) rigorously characterize the dependency of the SSM timescale on sequence length based on sequence autocorrelation; (2) find that with a proper timescale, allowing a zero real part for the eigenvalues of the SSM state matrix mitigates the curse of memory while still maintaining stability at initialization; (3) show that the imaginary part of the eigenvalues of the SSM state matrix determines the conditioning of SSM optimization problems, and uncover an approximation-estimation tradeoff when training SSMs with a specific class of target functions.

2624. ReCogLab: a framework testing relational reasoning & cognitive hypotheses on LLMs

链接：https://iclr.cc/virtual/2025/poster/27731 abstract： A fundamental part of human cognition is the ability to not only recall previous memories, but also reason across them to draw conclusions. In cognitive science and psychology, this is termed relational reasoning and a number of effects and biases have been observed in human cognition. Designing experiments to measure these reasoning effects is effortful and does not transfer easily to analyzing language model reasoning patterns. To make exploring language models on relational reasoning easier, we introduce ReCogLab – a generative framework for constructing reasoning examples. Unlike static datasets, our framework has a number of benefits that help us in our goal of flexible evaluation of LLMs. First, our framework allows us to control the difficulty and context-length of the problem, allowing us to scale with model capability and evaluate LLMs at a variety of scales. Second, the ability to change the configuration of a dataset dynamically allows us to probe models on different aspects and capabilities. Finally, the flexibility of our approach enables the recreation of classic cognitive science experiments and the systematic study of relational reasoning biases in language models. We demonstrate several such experiments and present our findings on a wide variety of open and closed-source language models. We release all data and code at https://github.com/google-deepmind/recoglab.

2625. GraphRouter: A Graph-based Router for LLM Selections

链接：https://iclr.cc/virtual/2025/poster/28926 abstract： The rapidly growing number and variety of Large Language Models (LLMs)present significant challenges in efficiently selecting the appropriate LLM fora given query, especially considering the trade-offs between performance andcomputational cost. Current LLM selection methods often struggle to generalizeacross new LLMs and different tasks because of their limited ability to leveragecontextual interactions among tasks, queries, and LLMs, as well as their depen-dence on a transductive learning framework. To address these shortcomings, weintroduce a novel inductive graph framework, named as GraphRouter, whichfully utilizes the contextual information among tasks, queries, and LLMs to en-hance the LLM selection process. GraphRouter constructs a heterogeneousgraph comprising task, query, and LLM nodes, with interactions represented asedges, which efficiently captures the contextual information between the query'srequirements and the LLM's capabilities. Through an innovative edge predictionmechanism, GraphRouter is able to predict attributes (the effect and cost ofLLM response) of potential edges, allowing for optimized recommendations thatadapt to both existing and newly introduced LLMs without requiring retraining.Comprehensive experiments across three distinct effect-cost weight scenarios haveshown that GraphRouter substantially surpasses existing routers, delivering aminimum performance improvement of 12.3%. In addition, it achieves enhancedgeneralization across new LLMs settings and supports diverse tasks with at least a9.5% boost in effect and a significant reduction in computational demands. Thiswork endeavors to apply a graph-based approach for the contextual and adaptiveselection of LLMs, offering insights for real-world applications.

2626. SOAP: Improving and Stabilizing Shampoo using Adam for Language Modeling

链接：https://iclr.cc/virtual/2025/poster/30183 abstract： There is growing evidence of the effectiveness of Shampoo, a higher-order preconditioning method, over Adam in deep learning optimization tasks. However, Shampoo's drawbacks include additional hyperparameters and computational overhead when compared to Adam, which only updates running averages of first- and second-moment quantities. This work establishes a formal connection between Shampoo (implemented with the 1/2 power)

and Adafactor --- a memory-efficient approximation of Adam --- showing that Shampoo is equivalent to running Adafactor in the eigenbasis of Shampoo's preconditioner. This insight leads to the design of a simpler and computationally efficient algorithm: ShampoO with Adam in the Preconditioner's eigenbasis (SOAP).With regards to improving Shampoo's computational efficiency, the most straightforward approach would be to simply compute Shampoo's eigendecomposition less frequently. Unfortunately, as our empirical results show, this leads to performance degradation that worsens with this frequency. SOAP mitigates this degradation by continually updating the running average of the second moment, just as Adam does, but in the current (slowly changing) coordinate basis. Furthermore, since SOAP is equivalent to running Adam in a rotated space, it introduces only one additional hyperparameter (the preconditioning frequency) compared to Adam. We empirically evaluate SOAP on language model pre-training with 360m and 660m sized models. In the large batch regime, SOAP reduces the number of iterations by over 40\% and wall clock time by over 35\% compared to AdamW, with approximately 20\% improvements in both metrics compared to Shampoo. An implementation of SOAP is available at https://github.com/nikhilvyas/SOAP.

2627. When is Task Vector Provably Effective for Model Editing? A Generalization Analysis of Nonlinear Transformers

链接：https://iclr.cc/virtual/2025/poster/27911 abstract： Task arithmetic refers to editing the pre-trained model by adding a weighted sum of task vectors, each of which is the weight update from the pre-trained model to fine-tuned models for certain tasks. This approach recently gained attention as a computationally efficient inference method for model editing, e.g., multi-task learning, forgetting, and out-of-domain generalization capabilities. However, the theoretical understanding of why task vectors can execute various conceptual operations remains limited, due to the highly non-convexity of training Transformer-based models. To the best of our knowledge, this paper provides the first theoretical characterization of the generalization guarantees of task vector methods on nonlinear Transformers. We consider a conceptual learning setting, where each task is a binary classification problem based on a discriminative pattern. We theoretically prove the effectiveness of task addition in simultaneously learning a set of irrelevant or aligned tasks, as well as the success of task negation in unlearning one task from irrelevant or contradictory tasks. Moreover, we prove the proper selection of linear coefficients for task arithmetic to achieve guaranteed generalization to out-of-domain tasks. All of our theoretical results hold for both dense-weight parameters and their low-rank approximations. Although established in a conceptual setting, our theoretical findings were validated on a practical machine unlearning task using the large language model Phi-1.5 (1.3B).

2628. Anti-Exposure Bias in Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29914 abstract： Diffusion models (DMs) have achieved record-breaking performance in image generation tasks.Nevertheless, in practice, the training-sampling discrepancy, caused by score estimation error and discretization error, limits the modeling ability of DMs, a phenomenon known as exposure bias.To alleviate such exposure bias and further improve the generative performance, we put forward a prompt learning framework built upon a lightweight prompt prediction model.Concretely, our model learns an anti-bias prompt for the generated sample at each sampling step, aiming to compensate for the exposure bias that arises.Following this design philosophy, our framework rectifies the sampling trajectory to match the training trajectory, thereby reducing the divergence between the target data distribution and the modeling distribution.To train the prompt prediction model, we simulate exposure bias by constructing training data and introduce a time-dependent weighting function for optimization.Empirical results on various DMs demonstrate the superiority of our prompt learning framework across three benchmark datasets.Importantly, the optimized prompt prediction model effectively improves image quality with only a 5\% increase in sampling overhead, which remains negligible.

2629. Easing Training Process of Rectified Flow Models Via Lengthening Inter-Path Distance

链接：https://iclr.cc/virtual/2025/poster/29645 abstract： Recent research pinpoints that different diffusion methods and architectures trained on the same dataset produce similar results for the same input noise. This property suggests that they have some preferable noises for a given sample. By visualizing the noise-sample pairs of rectified flow models and stable diffusion models in two-dimensional spaces, we observe that the preferable paths, connecting preferable noises to the corresponding samples, are better organized with significant fewer crossings comparing with the random paths, connecting random noises to training samples. In high-dimensional space, paths rarely intersect. The path crossings in two-dimensional spaces indicate the shorter inter-path distance in the corresponding high-dimensional spaces. Inspired by this observation, we propose the Distance-Aware Noise-Sample Matching (DANSM) method to lengthen the inter-path distance for speeding up the model training. DANSM is derived from rectified flow models, which allow using a closed-form formula to calculate the inter-path distance. To further simplify the optimization, we derive the relationship between inter-path distance and path length, and use the latter in the optimization surrogate. DANSM is evaluated on both image and latent spaces by rectified flow models and diffusion models. The experimental results show that DANSM can significantly improve the training speed by 30\% $\sim$ 40\%without sacrificing the generation quality.

2630. How Does Critical Batch Size Scale in Pre-training?

链接：https://iclr.cc/virtual/2025/poster/30121 abstract： Training large-scale models under given resources requires careful design of parallelism strategies. In particular, the efficiency notion of critical batch size (CBS), concerning the compromise

between time and compute, marks the threshold beyond which greater data parallelism leads to diminishing returns. To operationalize it, we propose a measure of CBS and pre-train a series of auto-regressive language models, ranging from 85 million to 1.2 billion parameters, on the C4 dataset. Through extensive hyper-parameter sweeps and careful control of factors such as batch size, momentum, and learning rate along with its scheduling, we systematically investigate the impact of scale on CBS. Then we fit scaling laws with respect to model and data sizes to decouple their effects. Overall, our results demonstrate that CBS scales primarily with data size rather than model size, a finding we justify theoretically through the analysis of infinite-width limits of neural networks and infinite-dimensional least squares regression. Of independent interest, we highlight the importance of common hyper-parameter choices and strategies for studying large-scale pre-training beyond fixed training durations.

2631. A New Perspective on Shampoo's Preconditioner

链接：https://iclr.cc/virtual/2025/poster/29069 abstract：

2632. Deconstructing What Makes a Good Optimizer for Autoregressive Language Models

链接：https://iclr.cc/virtual/2025/poster/27658 abstract： Training language models becomes increasingly expensive with scale, prompting numerous attempts to improve optimization efficiency. Despite these efforts, the Adam optimizer remains the most widely used, due to a prevailing view that it is the most effective approach. We aim to compare several optimization algorithms, including SGD, Adafactor, Adam, Lion, and Sophia in the context of autoregressive language modeling across a range of model sizes, hyperparameters, and architecture variants. Our findings indicate that, except for SGD, these algorithms all perform comparably both in their optimal performance and also in terms of how they fare across a wide range of hyperparameter choices. Our results suggest to practitioners that the choice of optimizer can be guided by practical considerations like memory constraints and ease of implementation, as no single algorithm emerged as a clear winner in terms of performance or stability to hyperparameter misspecification. Given our findings, we further dissect these approaches, examining two simplified versions of Adam: a) signed momentum (Signum) which we see recovers both the performance and hyperparameter stability of Adam and b) Adalayer, a layerwise variant of Adam which we introduce to study the impact on Adam's preconditioning for different layers of the network. Examining Adalayer leads us to the conclusion that, perhaps surprisingly, adaptivity on both the last layer and LayerNorm parameters in particular are necessary for retaining performance and stability to learning rate.

2633. Mixture of Parrots: Experts improve memorization more than reasoning

链接：https://iclr.cc/virtual/2025/poster/30684 abstract： The Mixture-of-Experts (MoE) architecture enables a significant increase in the total number of model parameters with minimal computational overhead. However, it is not clear what performance tradeoffs, if any, exist between MoEs and standard dense transformers.In this paper, we show that as we increase the number of experts (while fixing the number of active parameters), the memorization performance consistently increases while the reasoning capabilities saturate. We begin by analyzing the theoretical limitations of MoEs at reasoning. We prove that there exist graph problems that cannot be solved by any number of experts of a certain width; however, the same task can be easily solved by a dense model with a slightly larger width. On the other hand, we find that on memory-intensive tasks, MoEs can effectively leverage a small number of active parameters with a large number of experts to memorize the data. We empirically validate these findings on synthetic graph problems and memory-intensive closed book retrieval tasks. Lastly, we pre-train a series of MoEs and dense transformers and evaluate them on commonly used benchmarks in math and natural language. We find that increasing the number of experts helps solve knowledge-intensive tasks, but fails to yield the same benefits for reasoning tasks.

2634. Revisiting a Design Choice in Gradient Temporal Difference Learning

链接：https://iclr.cc/virtual/2025/poster/31094 abstract： Off-policy learning enables a reinforcement learning (RL) agent to reason counterfactually about policies that are not executed and is one of the most important ideas in RL. It, however, can lead to instability when combined with function approximation and bootstrapping, two arguably indispensable ingredients for large-scale reinforcement learning. This is the notorious deadly triad. The seminal work Sutton et al. (2008) pioneers Gradient Temporal Difference learning (GTD) as the first solution to the deadly triad, which has enjoyed massive success thereafter. During the derivation of GTD, some intermediate algorithm, called $A^\top$TD, was invented but soon deemed inferior. In this paper, we revisit this $A^\top$TD and prove that a variant of $A^\top$TD, called $A_t^\top$TD, is also an effective solution to the deadly triad. Furthermore, this $A_t^\top$TD only needs one set of parameters and one learning rate. By contrast, GTD has two sets of parameters and two learning rates, making it hard to tune in practice. We provide asymptotic analysis for $A^\top_t$TD and finite sample analysis for a variant of $A^\top_t$TD that additionally involves a projection operator. The convergence rate of this variant is on par with the canonical on-policy temporal difference learning.

2635. Towards Understanding Why FixMatch Generalizes Better Than Supervised Learning

链接：https://iclr.cc/virtual/2025/poster/31162 abstract： Semi-supervised learning (SSL), exemplified by FixMatch (Sohn et al., 2020), has shown significant generalization advantages over supervised learning (SL), particularly in the context of deep neural networks (DNNs). However, it is still unclear, from a theoretical standpoint, why FixMatch-like SSL algorithms generalize better than SL on DNNs. In this work, we present the first theoretical justification for the enhanced test accuracy observed in FixMatch-like SSL applied to DNNs by taking convolutional neural networks (CNNs) on classification tasks as an example. Our theoretical analysis reveals that the semantic feature learning processes in FixMatch and SL are rather different. In particular, FixMatch learns all the discriminative features of each semantic class, while SL only randomly captures a subset of features due to the well-known lottery ticket hypothesis. Furthermore, we show that our analysis framework can be applied to other FixMatch-like SSL methods, e.g., FlexMatch, FreeMatch, Dash, and SoftMatch. Inspired by our theoretical analysis, we develop an improved variant of FixMatch, termed Semantic-Aware FixMatch (SA-FixMatch). Experimental results corroborate our theoretical findings and the enhanced generalization capability of SA-FixMatch.

2636. Doubly Optimal Policy Evaluation for Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/30916 abstract： Policy evaluation estimates the performance of a policy by (1) collecting data from the environment and (2) processing raw data into a meaningful estimate. Due to the sequential nature of reinforcement learning, any improper data-collecting policy or data-processing method substantially deteriorates the variance of evaluation results over long time steps. Thus, policy evaluation often suffers from large variance and requires massive data to achieve the desired accuracy. In this work, we design an optimal combination of data-collecting policy and data-processing baseline. Theoretically, we prove our doubly optimal policy evaluation method is unbiased and guaranteed to have lower variance than previously best-performing methods. Empirically, compared with previous works, we show our method reduces variance substantially and achieves superior empirical performance.

2637. InsightBench: Evaluating Business Analytics Agents Through Multi-Step Insight Generation

链接：https://iclr.cc/virtual/2025/poster/29227 abstract： Data analytics is essential for extracting valuable insights from data that can assist organizations in making effective decisions. We introduce InsightBench, a benchmark dataset with three key features. First, it consists of 100 datasets representing diverse business use cases such as finance and incident management, each accompanied by a carefully curated set of insights planted in the datasets. Second, unlike existing benchmarks focusing on answering single queries, InsightBench evaluates agents based on their ability to perform end-to-end data analytics, including formulating questions, interpreting answers, and generating a summary of insights and actionable steps. Third, we conducted comprehensive quality assurance to ensure that each dataset in the benchmark had clear goals and included relevant and meaningful questions and analysis. Furthermore, we implement a two-way evaluation mechanism using LLaMA-3 as an effective, open-source evaluator to assess agents' ability to extract insights. We also propose AgentPoirot, our baseline data analysis agent capable of performing end-to-end data analytics. Our evaluation on InsightBench shows that AgentPoirot outperforms existing approaches (such as Pandas Agent) that focus on resolving single queries. We also compare the performance of open- and closed-source LLMs and various evaluation strategies. Overall, this benchmark serves as a testbed to motivate further development in comprehensive automated data analytics and can be accessed here: https://github.com/ServiceNow/insight-bench.

2638. Efficient Policy Evaluation with Safety Constraint for Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/30450 abstract： In reinforcement learning, classic on-policy evaluation methods often suffer from high variance and require massive online data to attain the desired accuracy. Previous studies attempt to reduce evaluation variance by searching for or designing proper behavior policies to collect data. However, these approaches ignore the safety of such behavior policies---the designed behavior policies have no safety guarantee and may lead to severe damage during online executions. In this paper, to address the challenge of reducing variance while ensuring safety simultaneously, we propose an optimal variance-minimizing behavior policy under safety constraints. Theoretically, while ensuring safety constraints, our evaluation method is unbiased and has lower variance than on-policy evaluation. Empirically, our method is the only existing method to achieve both substantial variance reduction and safety constraint satisfaction. Furthermore, we show our method is even superior to previous methods in both variance reduction and execution safety.

2639. Transformers Can Learn Temporal Difference Methods for In-Context Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29736 abstract： Traditionally, reinforcement learning (RL) agents learn to solve new tasks by updating their neural network parameters through interactions with the task environment. However, recent works demonstrate that some RL agents, after certain pretraining procedures, can learn to solve unseen new tasks without parameter updates, a phenomenon known as in-context reinforcement learning (ICRL). The empirical success of ICRL is widely attributed to the hypothesis that the forward pass of the pretrained agent neural network implements an RL algorithm. In this paper, we support this hypothesis by showing, both empirically and theoretically, that when a transformer is trained for policy evaluation tasks, it can discover and learn to implement temporal difference learning in its forward pass.

2640. LLaVA-Mini: Efficient Image and Video Large Multimodal Models with One Vision Token

链接：https://iclr.cc/virtual/2025/poster/29475 abstract：

2641. LLaMA-Omni: Seamless Speech Interaction with Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29748 abstract：

2642. Unleashing the Potential of Vision-Language Pre-Training for 3D Zero-Shot Lesion Segmentation via Mask-Attribute Alignment

链接：https://iclr.cc/virtual/2025/poster/32087 abstract：

2643. Rethinking the generalization of drug target affinity prediction algorithms via similarity aware evaluation

链接：https://iclr.cc/virtual/2025/poster/28654 abstract： Drug-target binding affinity prediction is a fundamental task for drug discovery. It has been extensively explored in literature and promising results are reported. However, in this paper, we demonstrate that the results may be misleading and cannot be well generalized to real practice. The core observation is that the canonical randomized split of a test set in conventional evaluation leaves the test set dominated by samples with high similarity to the training set. The performance of models is severely degraded on samples with lower similarity to the training set but the drawback is highly overlooked in current evaluation. As a result, the performance can hardly be trusted when the model meets low-similarity samples in real practice. To address this problem, we propose a framework of similarity aware evaluation in which a novel split methodology is proposed to adapt to any desired distribution. This is achieved by a formulation of optimization problems which are approximately and efficiently solved by gradient descent. We perform extensive experiments across five representative methods in four datasets for two typical target evaluations and compare them with various counterpart methods. Results demonstrate that the proposed split methodology can significantly better fit desired distributions and guide the development of models.

2644. MuirBench: A Comprehensive Benchmark for Robust Multi-image Understanding

链接：https://iclr.cc/virtual/2025/poster/29509 abstract： We introduce MuirBench, a comprehensive benchmark that focuses on robust multi-image understanding capabilities of multimodal LLMs. MuirBench consists of 12 diverse multi-image tasks (e.g., scene understanding, ordering) that involve 10 categories of multi-image relations (e.g., multiview, temporal relations). Comprising 11,264 images and 2,600 multiple-choice questions, MuirBench is created in a pairwise manner, where each standard instance is paired with an unanswerable variant that has minimal semantic differences, in order for a reliable assessment. Evaluated upon 20 recent multi-modal LLMs, our results reveal that even the best-performing models like GPT-4o and Gemini Pro find it challenging to solve MuirBench, achieving 68.0% and 49.3% in accuracy. Open-source multimodal LLMs trained on single images can hardly generalize to multi-image questions, hovering below 33.3% in accuracy. These results highlight the importance of MuirBench in encouraging the community to develop multimodal LLMs that can look beyond a single image, suggesting potential pathways for future improvements.

2645. Learning Shape-Independent Transformation via Spherical Representations for Category-Level Object Pose Estimation

链接：https://iclr.cc/virtual/2025/poster/30479 abstract： Category-level object pose estimation aims to determine the pose and size of novel objects in specific categories. Existing correspondence-based approaches typically adopt point-based representations to establish the correspondences between primitive observed points and normalized object coordinates. However, due to the inherent shape-dependence of canonical coordinates, these methods suffer from semantic incoherence across diverse object shapes. To resolve this issue, we innovatively leverage the sphere as a shared proxy shape of objects to learn shape-independent transformation via spherical representations. Based on this insight, we introduce a novel architecture called SpherePose, which yields precise correspondence prediction through three core designs. Firstly, We endow the point-wise feature extraction with SO(3)-invariance, which facilitates robust mapping between camera coordinate space and object coordinate space regardless of rotation transformation. Secondly, the spherical attention mechanism is designed to propagate and integrate features among spherical anchors from a comprehensive perspective, thus mitigating the interference of noise and incomplete point cloud. Lastly, a hyperbolic correspondence loss function is designed to distinguish subtle distinctions, which can promote the precision of correspondence prediction. Experimental results on CAMERA25, REAL275 and HouseCat6D benchmarks demonstrate the superior performance of our method, verifying the effectiveness of spherical representations and

architectural innovations.

2646. Improving Deep Regression with Tightness

链接：https://iclr.cc/virtual/2025/poster/28969 abstract： For deep regression, preserving the ordinality of the targets with respect to the feature representation improves performance across various tasks. However, a theoretical explanation for the benefits of ordinality is still lacking. This work reveals that preserving ordinality reduces the conditional entropy $H(Z|Y)$ of representation Z conditional on the target Y. However, our findings reveal that typical regression losses fail to sufficiently reduce $H(Z|Y)$, despite its crucial role in generalization performance. With this motivation, we introduce an optimal transport-based regularizer to preserve the similarity relationships of targets in the feature space to reduce $H(Z|Y)$. Additionally, we introduce a simple yet efficient strategy of duplicating the regressor targets, also with the aim of reducing $H(Z|Y)$. Experiments on three real-world regression tasks verify the effectiveness of our strategies to improve deep regression. Code: https://github.com/needylove/Regression_tightness

2647. RainbowPO: A Unified Framework for Combining Improvements in Preference Optimization

链接：https://iclr.cc/virtual/2025/poster/28018 abstract： Recently, numerous preference optimization algorithms have been introduced as extensions to the Direct Preference Optimization (DPO) family. While these methods have successfully aligned models with human preferences, there is a lack of understanding regarding the contributions of their additional components. Moreover, fair and consistent comparisons are scarce, making it difficult to discern which components genuinely enhance downstream performance. In this work, we propose RainbowPO, a unified framework that demystifies the effectiveness of existing DPO methods by categorizing their key components into seven broad directions. We integrate these components into a single cohesive objective, enhancing the performance of each individual element. Through extensive experiments, we demonstrate that RainbowPO outperforms existing DPO variants. Additionally, we provide insights to guide researchers in developing new DPO methods and assist practitioners in their implementations.

2648. Generative Verifiers: Reward Modeling as Next-Token Prediction

链接：https://iclr.cc/virtual/2025/poster/30506 abstract： Verifiers or reward models are often used to enhance the reasoning performance of large language models (LLMs). A common approach is the Best-of-N method, where N candidate solutions generated by the LLM are ranked by a verifier, and the best one is selected. While LLM-based verifiers are typically trained as discriminative classifiers to score solutions, they do not utilize the text generation capabilities of pretrained LLMs. To overcome this limitation, we instead propose training verifiers using the ubiquitous next-token prediction objective, jointly on verification and solution generation. Compared to standard verifiers, such generative verifiers (GenRM) can benefit from several advantages of LLMs: they integrate seamlessly with instruction tuning, enable chain-of-thought reasoning, and can utilize additional test-time compute via majority voting for better verification. We demonstrate that GenRM outperforms discriminative, DPO verifiers, and LLM-as-a-Judge, resulting in large performance gains with Best-of-N, namely 5% → 45.3% on algorithmic tasks, 73% → 93.4% on GSM8K, and 28% →44.6% on easy-to-hard generalization on MATH. Furthermore, we find that training GenRM with synthetic verification rationales is sufficient to pick out subtle errors on math problems. Finally, we demonstrate that generative verifiers scale favorably with model size and inference-time compute.

2649. GSE: Group-wise Sparse and Explainable Adversarial Attacks

链接：https://iclr.cc/virtual/2025/poster/29010 abstract： Sparse adversarial attacks fool deep neural networks (DNNs) through minimal pixel perturbations, often regularized by the ℓ_0 norm. Recent efforts have replaced this norm with a structural sparsity regularizer, such as the nuclear group norm, to craft group-wise sparse adversarial attacks. The resulting perturbations are thus explainable and hold significant practical relevance, shedding light on an even greater vulnerability of DNNs. However, crafting such attacks poses an optimization challenge, as it involves computing norms for groups of pixels within a non-convex objective. We address this by presenting a two-phase algorithm that generates group-wise sparse attacks within semantically meaningful areas of an image. Initially, we optimize a quasinorm adversarial loss using the $1/2$-quasinorm proximal operator tailored for non-convex programming. Subsequently, the algorithm transitions to a projected Nesterov's accelerated gradient descent with 2-norm regularization applied to perturbation magnitudes. Rigorous evaluations on CIFAR-10 and ImageNet datasets demonstrate a remarkable increase in group-wise sparsity, e.g., 50.9% on CIFAR-10 and 38.4% on ImageNet (average case, targeted attack). This performance improvement is accompanied by significantly faster computation times, improved interpretability, and a 100% attack success rate.

2650. Understanding Methods for Scalable MCTS

链接：https://iclr.cc/virtual/2025/poster/31350 abstract： Monte Carlo Tree Search (MCTS) is a versatile algorithm widely used for intelligent decision-making in complex, high-dimensional environments. While MCTS inherently improves with more compute, real-world applications often demand rapid decision-making under strict inference-time constraints. This blog post explores scalable parallelization strategies for MCTS, covering classical methods (leaf, root, and tree parallelism) and advanced distributed approaches—including virtual loss, transposition-driven scheduling, and distributed depth-first scheduling. By examining the practical trade-offs and performance implications of each method, we identify effective techniques for achieving

high-throughput, low-latency planning—critical for applications like autonomous vehicles, emergency response systems, and real-time trading.

2651. Dynamic Multimodal Evaluation with Flexible Complexity by Vision-Language Bootstrapping

链接：https://iclr.cc/virtual/2025/poster/29328 abstract： Large Vision-Language Models (LVLMs) have demonstrated remarkable capabilities across multimodal tasks such as visual perception and reasoning, leading to good performance on various multimodal evaluation benchmarks. However, these benchmarks keep a static nature and overlap with the pre-training data, resulting in fixed complexity constraints and data contamination issues. This raises the concern regarding the validity of the evaluation. To address these two challenges, we introduce a dynamic multimodal evaluation protocol called Vision-Language Bootstrapping (VLB). VLB provides a robust and comprehensive assessment for LVLMs with reduced data contamination and flexible complexity. To this end, VLB dynamically generates new visual question-answering samples through a multimodal bootstrapping module that modifies both images and language, while ensuring that newly generated samples remain consistent with the original ones by a judge module. By composing various bootstrapping strategies, VLB offers dynamic variants of existing benchmarks with diverse complexities, enabling the evaluation to co-evolve with the ever-evolving capabilities of LVLMs. Extensive experimental results across multiple benchmarks, including SEEDBench, MMBench, and MME, show that VLB significantly reduces data contamination and exposes performance limitations of LVLMs.

2652. Neural Sampling from Boltzmann Densities: Fisher-Rao Curves in the Wasserstein Geometry

链接：https://iclr.cc/virtual/2025/poster/29539 abstract：

2653. eQMARL: Entangled Quantum Multi-Agent Reinforcement Learning for Distributed Cooperation over Quantum Channels

链接：https://iclr.cc/virtual/2025/poster/29050 abstract：

2654. AnoLLM: Large Language Models for Tabular Anomaly Detection

链接：https://iclr.cc/virtual/2025/poster/30820 abstract：

2655. The impact of allocation strategies in subset learning on the expressive power of neural networks

链接：https://iclr.cc/virtual/2025/poster/27957 abstract：

2656. Fine-tuning can Help Detect Pretraining Data from Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29320 abstract： In the era of large language models (LLMs), detecting pretraining data has been increasingly important due to concerns about fair evaluation and ethical risks. Current methods differentiate members and non-members by designing scoring functions, like Perplexity and Min-k%. However, the diversity and complexity of training data magnifies the difficulty of distinguishing, leading to suboptimal performance in detecting pretraining data. In this paper, we first explore the benefits of unseen data, which can be easily collected after the release of the LLM. We find that the perplexities of LLMs shift differently for members and non-members, after fine-tuning with a small amount of previously unseen data. In light of this, we introduce a novel and effective method termed Fine-tuned Score Deviation (FSD), which improves the performance of current scoring functions for pretraining data detection. In particular, we propose to measure the deviation distance of current scores after fine-tuning on a small amount of unseen data within the same domain. In effect, using a few unseen data can largely decrease the scores of all non-members, leading to a larger deviation distance than members. Extensive experiments demonstrate the effectiveness of our method, significantly improving the AUC score on common benchmark datasets across various models.

2657. Boost Self-Supervised Dataset Distillation via Parameterization, Predefined Augmentation, and Approximation

链接：https://iclr.cc/virtual/2025/poster/31137 abstract： Although larger datasets are crucial for training large deep models, the rapid growth of dataset size has brought a significant challenge in terms of considerable training costs, which even results in prohibitive computational expenses. Dataset Distillation becomes a popular technique recently to reduce the dataset size via learning a highly compact set of representative exemplars, where the model trained with these exemplars ideally should have

comparable performance with respect to the one trained with the full dataset. While most of existing works upon dataset distillation focus on supervised datasets, \todo{we instead aim to distill images and their self-supervisedly trained representations into a distilled set. This procedure, named as Self-Supervised Dataset Distillation, effectively extracts rich information from real datasets, yielding the distilled sets with enhanced cross-architecture generalizability.} Particularly, in order to preserve the key characteristics of original dataset more faithfully and compactly, several novel techniques are proposed: 1) we introduce an innovative parameterization upon images and representations via distinct low-dimensional bases, where the base selection for parameterization is experimentally shown to play a crucial role; 2) we tackle the instability induced by the randomness of data augmentation -- a key component in self-supervised learning but being underestimated in the prior work of self-supervised dataset distillation -- by utilizing predetermined augmentations; 3) we further leverage a lightweight network to model the connections among the representations of augmented views from the same image, leading to more compact pairs of distillation. Extensive experiments conducted on various datasets validate the superiority of our approach in terms of distillation efficiency, cross-architecture generalization, and transfer learning performance.

2658. Exploring Learning Complexity for Efficient Downstream Dataset Pruning

链接：https://iclr.cc/virtual/2025/poster/30346 abstract： The ever-increasing fine-tuning cost of large-scale pre-trained models gives rise to the importance of dataset pruning, which aims to reduce dataset size while maintaining task performance.However, existing dataset pruning methods require training on the entire dataset, which is impractical for large-scale pre-trained models.In this paper, we propose a straightforward, novel, and training-free hardness score named Distorting-based Learning Complexity (DLC), to identify informative images and instructions from the downstream dataset efficiently.Our method is motivated by the observation that easy samples learned faster can also be learned with fewer parameters.Specifically, we define the Learning Complexity to quantify sample hardness and utilize a lightweight weights masking process for fast estimation, instead of the costly SGD optimization.Based on DLC, we further design a flexible under-sampling strategy with randomness (dubbed FlexRand), replacing the top-K strategy, to alleviate the severe subset distribution shift.Extensive experiments with downstream image and instructions dataset pruning benchmarks demonstrate the effectiveness and efficiency of the proposed approach.In the images pruning benchmark, DLC significantly reduces the pruning time by 35$\times$ while establishing state-of-the-art performance with FlexRand.

2659. REvolve: Reward Evolution with Large Language Models using Human Feedback

链接：https://iclr.cc/virtual/2025/poster/29060 abstract： Designing effective reward functions is crucial to training reinforcement learning (RL) algorithms. However, this design is non-trivial, even for domain experts, due to the subjective nature of certain tasks that are hard to quantify explicitly. In recent works, large language models (LLMs) have been used for reward generation from natural language task descriptions, leveraging their extensive instruction tuning and commonsense understanding of human behavior. In this work, we hypothesize that LLMs, guided by human feedback, can be used to formulate reward functions that reflect human implicit knowledge. We study this in three challenging settings -- autonomous driving, humanoid locomotion, and dexterous manipulation -- wherein notions of ``good" behavior are tacit and hard to quantify. To this end, we introduce REvolve, a truly evolutionary framework that uses LLMs for reward design in RL. REvolve generates and refines reward functions by utilizing human feedback to guide the evolution process, effectively translating implicit human knowledge into explicit reward functions for training (deep) RL agents. Experimentally, we demonstrate that agents trained on REvolve-designed rewards outperform other state-of-the-art baselines.

2660. UNIP: Rethinking Pre-trained Attention Patterns for Infrared Semantic Segmentation

链接：https://iclr.cc/virtual/2025/poster/32078 abstract： Pre-training techniques significantly enhance the performance of semantic segmentation tasks with limited training data. However, the efficacy under a large domain gap between pre-training (e.g. RGB) and fine-tuning (e.g. infrared) remains underexplored. In this study, we first benchmark the infrared semantic segmentation performance of various pre-training methods and reveal several phenomena distinct from the RGB domain. Next, our layerwise analysis of pre-trained attention maps uncovers that: (1) There are three typical attention patterns (local, hybrid, and global); (2) Pre-training tasks notably influence pattern distribution across layers; (3) The hybrid pattern is crucial for semantic segmentation as it attends to both nearby and foreground elements; (4) The texture bias impedes model generalization in infrared tasks. Building on these insights, we propose UNIP, a UNified Infrared Pre-training framework, to enhance the pre-trained model performance. This framework uses the hybrid-attention distillation NMI-HAD as the pre-training target, a large-scale mixed dataset InfMix for pre-training, and a last-layer feature pyramid network LL-FPN for fine-tuning. Experimental results show that UNIP outperforms various pre-training methods by up to 13.5% in average mIoU on three infrared segmentation tasks, evaluated using fine-tuning and linear probing metrics. UNIP-S achieves performance on par with MAE-L while requiring only 1/10 of the computational cost. Furthermore, with fewer parameters, UNIP significantly surpasses state-of-the-art (SOTA) infrared or RGB segmentation methods and demonstrates the broad potential for application in other modalities, such as RGB and depth. Our code is available at https://github.com/casiatao/UNIP.

2661. Wavelet Diffusion Neural Operator

链接：https://iclr.cc/virtual/2025/poster/30342 abstract： Simulating and controlling physical systems described by partial differential equations (PDEs) are crucial tasks across science and engineering. Recently, diffusion generative models have emerged as a competitive class of methods for these tasks due to their ability to capture long-term dependencies and model high-dimensional states. However, diffusion models typically struggle with handling system states with abrupt changes and generalizing to higher resolutions. In this work, we propose Wavelet Diffusion Neural Operator (WDNO), a novel PDE simulation and control framework that enhances the handling of these complexities. WDNO comprises two key innovations. Firstly, WDNO performs diffusion-based generative modeling in the wavelet domain for the entire trajectory to handle abrupt changes and long-term dependencies effectively. Secondly, to address the issue of poor generalization across different resolutions, which is one of the fundamental tasks in modeling physical systems, we introduce multi-resolution training. We validate WDNO on five physical systems, including 1D advection equation, three challenging physical systems with abrupt changes (1D Burgers' equation, 1D compressible Navier-Stokes equation and 2D incompressible fluid), and a real-world dataset ERA5, which demonstrates superior performance on both simulation and control tasks over state-of-the-art methods, with significant improvements in long-term and detail prediction accuracy. Remarkably, in the challenging context of the 2D high-dimensional and indirect control task aimed at reducing smoke leakage, WDNO reduces the leakage by 33.2% compared to the second-best baseline.

2662. CL-DiffPhyCon: Closed-loop Diffusion Control of Complex Physical Systems

链接：https://iclr.cc/virtual/2025/poster/29738 abstract： The control problems of complex physical systems have broad applications in science and engineering. Previous studies have shown that generative control methods based on diffusion models offer significant advantages for solving these problems. However, existing generative control approaches face challenges in both performance and efficiency when extended to the closed-loop setting, which is essential for effective control. In this paper, we propose an efficient Closed-Loop Diffusion method for Physical systems Control (CL-DiffPhyCon). By employing an asynchronous denoising framework for different physical time steps, CL-DiffPhyCon generates control signals conditioned on real-time feedback from the system with significantly reduced computational cost during sampling. Additionally, the control process could be further accelerated by incorporating fast sampling techniques, such as DDIM. We evaluate CL-DiffPhyCon on two tasks: 1D Burgers' equation control and 2D incompressible fluid control. The results demonstrate that CL-DiffPhyCon achieves superior control performance with significant improvements in sampling efficiency. The code can be found at https://github.com/AI4Science-WestlakeU/CL_DiffPhyCon.

2663. Metamizer: A Versatile Neural Optimizer for Fast and Accurate Physics Simulations

链接：https://iclr.cc/virtual/2025/poster/30915 abstract： Efficient physics simulations are essential for numerous applications, ranging from realistic cloth animations in video games, to analyzing pollutant dispersion in environmental sciences, to calculating vehicle drag coefficients in engineering applications. Unfortunately, analytical solutions to the underlying physical equations are rarely available, and numerical solutions are computationally demanding. Latest developments in the field of physics-based Deep Learning have led to promising efficiency gains but still suffer from limited generalization capabilities across multiple different PDEs. Thus, in this work, we introduce Metamizer, a novel neural optimizer that iteratively solves a wide range of physical systems without retraining by minimizing a physics-based loss function. To this end, our approach leverages a scale-invariant architecture that enhances gradient descent updates to accelerate convergence. Since the neural network itself acts as an optimizer, training this neural optimizer falls into the category of meta-optimization approaches. We demonstrate that Metamizer achieves high accuracy across multiple PDEs after training on the Laplace, advection-diffusion and incompressible Navier-Stokes equation as well as on cloth simulations. Remarkably, the model also generalizes to PDEs that were not covered during training such as the Poisson, wave and Burgers equation.

2664. Long-tailed Adversarial Training with Self-Distillation

链接：https://iclr.cc/virtual/2025/poster/27919 abstract： Adversarial training significantly enhances adversarial robustness, yet superior performance is predominantly achieved on balanced datasets. Addressing adversarial robustness in the context of unbalanced or long-tailed distributions is considerably more challenging, mainly due to the scarcity of tail data instances. Previous research on adversarial robustness within long-tailed distributions has primarily focused on combining traditional long-tailed natural training with existing adversarial robustness methods. In this study, we provide an in-depth analysis for the challenge that adversarial training struggles to achieve high performance on tail classes in long-tailed distributions. Furthermore, we propose a simple yet effective solution to advance adversarial robustness on long-tailed distributions through a novel self-distillation technique. Specifically, this approach leverages a balanced self-teacher model, which is trained using a balanced dataset sampled from the original long-tailed dataset.Our extensive experiments demonstrate state-of-the-art performance in both clean and robust accuracy for long-tailed adversarial robustness, with significant improvements in tail class performance on various datasets.We improve the accuracy against PGD attacks for tail classes by 20.3, 7.1, and 3.8 percentage points on CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively, while achieving the highest robust accuracy.

2665. TaskGalaxy: Scaling Multi-modal Instruction Fine-tuning with Tens of Thousands Vision Task Types

链接：https://iclr.cc/virtual/2025/poster/32097 abstract： Multimodal visual language models are gaining prominence in open-world applications, driven by advancements in model architectures, training techniques, and high-quality data. However, their performance is often limited by insufficient task-specific data, leading to poor generalization and biased outputs. Existing efforts to increase task diversity in fine-tuning datasets are hindered by the labor-intensive process of manual task labeling, which typically produces only a few hundred task types. To address this, we propose TaskGalaxy, a large-scale multimodal instruction fine-tuning dataset comprising 19,227 hierarchical task types and 413,648 samples. TaskGalaxy utilizes GPT-4o to enrich task diversity by expanding from a small set of manually defined tasks, with CLIP and GPT-4o filtering those that best match open-source images, and generating relevant question-answer pairs. Multiple models are employed to ensure sample quality. This automated process enhances both task diversity and data quality, reducing manual intervention. Incorporating TaskGalaxy into LLaVA-v1.5 and InternVL-Chat-v1.0 models shows substantial performance improvements across 16 benchmarks, demonstrating the critical importance of task diversity. TaskGalaxy is publicly released at https://github.com/Kwai-YuanQi/TaskGalaxy.

2666. VCR: A Task for Pixel-Level Complex Reasoning in Vision Language Models via Restoring Occluded Text

链接：https://iclr.cc/virtual/2025/poster/28154 abstract： We introduce Visual Caption Restoration (VCR), a novel vision-language task that challenges models to accurately restore partially obscured texts using pixel-level hints within images through complex reasoning. This task stems from the observation that text embedded in images intrinsically differs from common visual elements and text due to the need to align the modalities of vision, text, and text embedded in images. While many works incorporate text into images for visual question answering, they mostly rely on OCR or masked language modeling, reducing the task to text-based processing. However, text-based processing becomes ineffective in VCR as accurate text restoration depends on the combined information from provided images, context, and subtle cues from the tiny, exposed areas of masked texts. We develop a pipeline to generate synthetic images for the VCR task using image-caption pairs, with adjustable caption visibility to control the task difficulty. With this pipeline, we construct VCR-WIKI for VCR using Wikipedia images with captions, including 2.11M English and 346K Chinese training entities, plus 5K validation and 5K test entities in both languages, each in easy and hard configurations. We also make a hidden test set, VCR-HIDDEN, to avoid potential overfitting on VCR-WIKI. Our results reveal that current vision-language models significantly lag behind human performance in the VCR task, and merely fine-tuning the models on our dataset does not lead to notable improvements. We release VCR-WIKI and the data construction code to facilitate future research.

2667. MallowsPO: Fine-Tune Your LLM with Preference Dispersions

链接：https://iclr.cc/virtual/2025/poster/29005 abstract： Direct Preference Optimization (DPO) has recently emerged as a popular approach to improve reinforcement learning from human feedback (RLHF), leading to better techniques to fine-tune large language models (LLM). A weakness of DPO, however, lies in its lack of capability to characterize the diversity of human preferences. Inspired by Mallows' theory of preference ranking, we develop in this paper a new approach, the MallowsPO. A distinct feature of this approach is a dispersion index, which reflects the dispersion of human preference to prompts. We show that existing DPO models can be reduced to special cases of this dispersion index, thus unified with MallowsPO. More importantly, we demonstrate empirically how to use this dispersion index to enhance the performance of DPO in a broad array of benchmark tasks, from synthetic bandit selection to controllable generation and dialogues, while maintaining great generalization capabilities. MallowsPO is also compatible with other SOTA offline preference optimization methods, boosting nearly 2\% extra LC win rate when used as a plugin for fine-tuning Llama3-Instruct.

2668. MoS: Unleashing Parameter Efficiency of Low-Rank Adaptation with Mixture of Shards

链接：https://iclr.cc/virtual/2025/poster/31177 abstract： The rapid scaling of large language models necessitates more lightweight finetuning methods to reduce the explosive GPU memory overhead when numerous customized models are served simultaneously.Targeting more parameter-efficient low-rank adaptation (LoRA), parameter sharing presents a promising solution. Empirically, our research into high-level sharing principles highlights the indispensable role of differentiation in reversing the detrimental effects of pure sharing.Guided by this finding, we propose Mixture of Shards (MoS), incorporating both inter-layer and intra-layer sharing schemes, and integrating four nearly cost-free differentiation strategies, namely subset selection, pair dissociation, vector sharding, and shard privatization. Briefly, it selects a designated number of shards from global pools with a Mixture-of-Experts (MoE)-like routing mechanism before sequentially concatenating them to low-rank matrices.Hence, it retains all the advantages of LoRA while offering enhanced parameter efficiency, and effectively circumvents the drawbacks of peer parameter-sharing methods.Our empirical experiments demonstrate approximately $8\times$ parameter savings in a standard LoRA setting. The ablation study confirms the significance of each component.Our insights into parameter sharing and MoS method may illuminate future developments of more parameter-efficient finetuning methods.The code is officially available at https://github.com/Forence1999/MoS.

2669. MAP: Low-compute Model Merging with Amortized Pareto Fronts via Quadratic Approximation

链接：https://iclr.cc/virtual/2025/poster/31176 abstract： Model merging has emerged as an effective approach to combining multiple single-task models into a multitask model. This process typically involves computing a weighted average of the model parameters without additional training. Existing model-merging methods focus on improving average task accuracy. However, interference and conflicts between the objectives of different tasks can lead to trade-offs during the merging process. In real-world applications, a set of solutions with various trade-offs can be more informative, helping practitioners make decisions based on diverse preferences. In this paper, we introduce a novel and low-compute algorithm, Model Merging with Amortized Pareto Front (MAP). MAP efficiently identifies a Pareto set of scaling coefficients for merging multiple models, reflecting the trade-offs involved. It amortizes the substantial computational cost of evaluations needed to estimate the Pareto front by using quadratic approximation surrogate models derived from a preselected set of scaling coefficients. Experimental results on vision and natural language processing tasks demonstrate that MAP can accurately identify the Pareto front, providing practitioners with flexible solutions to balance competing task objectives. We also introduce Bayesian MAP for scenarios with a relatively low number of tasks and Nested MAP for situations with a high number of tasks, further reducing the computational cost of evaluation.

2670. OccProphet: Pushing the Efficiency Frontier of Camera-Only 4D Occupancy Forecasting with an Observer-Forecaster-Refiner Framework

链接：https://iclr.cc/virtual/2025/poster/27927 abstract： Predicting variations in complex traffic environments is crucial for the safety of autonomous driving. Recent advancements in occupancy forecasting have enabled forecasting future 3D occupied status in driving environments by observing historical 2D images. However, high computational demands make occupancy forecasting less efficient during training and inference stages, hindering its feasibility for deployment on edge agents. In this paper, we propose a novel framework, \textit{i.e.}, OccProphet, to efficiently and effectively learn occupancy forecasting with significantly lower computational requirements while improving forecasting accuracy. OccProphet comprises three lightweight components: Observer, Forecaster, and Refiner. The Observer extracts spatio-temporal features from 3D multi-frame voxels using the proposed Efficient 4D Aggregation with Tripling-Attention Fusion, while the Forecaster and Refiner conditionally predict and refine future occupancy inferences. Experimental results on nuScenes, Lyft-Level5, and nuScenes-Occupancy datasets demonstrate that OccProphet is both training- and inference-friendly. OccProphet reduces 58\%$\sim$78\% of the computational cost with a 2.6$\times$ speedup compared with the state-of-the-art Cam4DOcc. Moreover, it achieves 4\%$\sim$18\% relatively higher forecasting accuracy. Code and models are publicly available at https://github.com/JLChen-C/OccProphet.

2671. Painting with Words: Elevating Detailed Image Captioning with Benchmark and Alignment Learning

链接：https://iclr.cc/virtual/2025/poster/30910 abstract： Image captioning has long been a pivotal task in visual understanding, with recent advancements in vision-language models (VLMs) significantly enhancing the ability to generate detailed image captions. However, the evaluation of detailed image captioning remains underexplored due to outdated evaluation metrics and coarse annotations. In this paper, we introduce DeCapBench along with a novel metric, DCScore, specifically designed for detailed captioning tasks. DCScore evaluates hallucinations and fine-grained comprehensiveness by deconstructing responses into the smallest self-sufficient units, termed primitive information units, and assessing them individually. Our evaluation shows that DCScore aligns more closely with human judgment than other rule-based or model-based metrics. Concurrently, DeCapBench exhibits a high correlation with VLM arena results on descriptive tasks, surpassing existing benchmarks for vision-language models. Additionally, we present an automatic fine-grained feedback collection method, FeedQuill, for preference optimization based on our advanced metric, demonstrating robust generalization capabilities across auto-generated preference data. Extensive experiments on multiple VLMs demonstrate that our method not only significantly reduces hallucinations but also enhances performance across various benchmarks, achieving superior detail captioning performance while surpassing GPT-4o.

2672. Nonlinear Sequence Embedding by Monotone Variational Inequality

链接：https://iclr.cc/virtual/2025/poster/29488 abstract： In the wild, we often encounter collections of sequential data such as electrocardiograms, motion capture, genomes, and natural language, and sequences may be multichannel or symbolic with nonlinear dynamics. We introduce a method to learn low-dimensional representations of nonlinear sequence and time-series data without supervision which has provable recovery guarantees. The learned representation can be used for downstream machine-learning tasks such as clustering and classification. The method assumes that the observed sequences arise from a common domain, with each sequence following its own autoregressive model, and these models are related through low-rank regularization. We cast the problem as a convex matrix parameter recovery problem using monotone variational inequalities (VIs) and encode the common domain assumption via low-rank constraint across the learned representations, which can learn a subspace approximately spanning the entire domain as well as faithful representations for the dynamics of each individual sequence incorporating the domain information in totality. We show the competitive performance of our method on real-world time-series data with baselines and demonstrate its effectiveness for symbolic text modeling and RNA sequence clustering.

2673. Kernel-based Optimally Weighted Conformal Time-Series Prediction

链接：https://iclr.cc/virtual/2025/poster/28361 abstract： Conformal prediction has been a popular distribution-free framework for uncertainty quantification. In this work, we present a novel conformal prediction method for time-series, which we call Kernel-

based Optimally Weighted Conformal Prediction Intervals ($\texttt{KOWCPI}$). Specifically, $\texttt{KOWCPI}$ adapts the classic Reweighted Nadaraya-Watson (RNW) estimator for quantile regression on dependent data and learns optimal data-adaptive weights. Theoretically, we tackle the challenge of establishing a conditional coverage guarantee for non-exchangeable data under strong mixing conditions on the non-conformity scores. We demonstrate the superior performance of $\texttt{KOWCPI}$ on real time-series against state-of-the-art methods, where $\texttt{KOWCPI}$ achieves narrower confidence intervals without losing coverage.

2674. Agree to Disagree: Demystifying Homogeneous Deep Ensembles through Distributional Equivalence

链接：https://iclr.cc/virtual/2025/poster/29301 abstract： Deep ensembles improve the performance of the models by taking the average predictions of a group of ensemble members. However, the origin of these capabilities remains a mystery and deep ensembles are used as a reliable “black box” to improve the performance. Existing studies typically attribute such improvement to Jensen gaps of the deep ensemble method, where the loss of the mean does not exceed the mean of the loss for any convex loss metric. In this work, we demonstrate that Jensen’s inequality is not responsible for the effectiveness of deep ensembles, and convexity is not a necessary condition. Instead, Jensen Gap focuses on the “average loss” of individual models, which provides no practical meaning. Thus it fails to explain the core phenomena of deep ensembles such as the superiority to any single ensemble member, the decreasing loss with the number of ensemble members, etc. Regarding this mystery, we provide theoretical analysis and comprehensive empirical results from a statistical perspective that reveal the true mechanism of deep ensembles. Our results highlight that deep ensembles originate from the homogeneous output distribution across all ensemble members. Specifically, the predictions of homogeneous models (Abe et al., 2022b) have the distributional equivalence property – Although the predictions of independent ensemble members are point-wise different, they form an identical distribution. Such agreement and disagreement contribute to deep ensembles’ “magical power”. Based on this discovery, we provide rigorous proof of the effectiveness of deep ensembles and analytically quantify the extent to which ensembles improve performance. The derivations not only theoretically quantify the effectiveness of deep ensembles for the first time, but also enable estimation schemes that foresee the performance of ensembles with different capacities. Furthermore, different from existing studies, our results also point out that deep ensembles work in a different mechanism from model scaling a single model, even though significant correlations between them have been observed.

2675. Quantum (Inspired) D^2-sampling with Applications

链接：https://iclr.cc/virtual/2025/poster/28068 abstract： D^2-sampling is a fundamental component of sampling-based clustering algorithms such as k-means++. Given a dataset $V \subset \mathbb{R}^d$ with N points and a center set $C \subset \mathbb{R}^d$, D^2-sampling refers to picking a point from V where the sampling probability of a point is proportional to its squared distance from the nearest center in C.The popular k-means++ algorithm is simply a k-round D^2-sampling process, which runs in $O(Nkd)$ time and gives $O(\log{k})$-approximation in expectation for the k-means problem.In this work, we give a quantum algorithm for (approximate) D^2-sampling in the QRAM model that results in a quantum implementation of k-means++ with a running time $\tilde{O}(\zeta^2 k^2)$. Here ζ is the aspect ratio (i.e., largest to smallest interpoint distance) and $\tilde{O}$ hides polylogarithmic factors in N, d, k.It can be shown through a robust approximation analysis of k-means++ that the quantum version preserves its $O(\log{k})$ approximation guarantee.Further, we show that our quantum algorithm for D^2-sampling can be dequantized using the sample-query access model of Tang (PhD Thesis, Ewin Tang, University of Washington, 2023). This results in a fast quantum-inspired classical implementation of k-means++, which we call QI-k-means++, with a running time $O(Nd) + \tilde{O}(\zeta^2k^2d)$, where the $O(Nd)$ term is for setting up the sample-query access data structure.Experimental investigations show promising results for QI-k-means++ on large datasets with bounded aspect ratio.Finally, we use our quantum D^2-sampling with the known $ D^2$-sampling-based classical approximation scheme to obtain the first quantum approximation scheme for the k-means problem with polylogarithmic running time dependence on N.

2676. Inspection and Control of Self-Generated-Text Recognition Ability in Llama3-8b-Instruct

链接：https://iclr.cc/virtual/2025/poster/27838 abstract： It has been reported that LLMs can recognize their own writing. As this has potential implications for AI safety, yet is relatively understudied, we investigate the phenomenon, seeking to establish: whether it robustly occurs at the behavioral level, how the observed behavior is achieved, and whether it can be controlled. First, we find that the Llama3-8b–Instruct chat model - but not the base Llama3-8b model - can reliably distinguish its own outputs from those of humans, and present evidence that the chat model is likely using its experience with its own outputs, acquired during post-training, to succeed at the writing recognition task. Second, we identify a vector in the residual stream of the model that is differentially activated when the model makes a correct self-written-text recognition judgment, show that the vector activates in response to information relevant to self-authorship, present evidence that the vector is related to the concept of ``self'' in the model, and demonstrate that the vector is causally related to the model’s ability to perceive and assert self-authorship. Finally, we show that the vector can be used to control both the model’s behavior and its perception, steering the model to claim or disclaim authorship by applying the vector to the model’s output as it generates it, and steering the model to believe or disbelieve it wrote arbitrary texts by applying the vector to them as the model reads them.

2677. Concept Bottleneck Language Models For Protein Design

链接：https://iclr.cc/virtual/2025/poster/29243 abstract： We introduce Concept Bottleneck Protein Language Models (CB-pLM), a generative masked language model with a layer where each neuron corresponds to an interpretable concept. Our architecture offers three key benefits: i) Control: We can intervene on concept values to precisely control the properties of generated proteins, achieving a 3$\times$ larger change in desired concept values compared to baselines. ii) Interpretability: A linear mapping between concept values and predicted tokens allows transparent analysis of the model's decision-making process. iii) Debugging: This transparency facilitates easy debugging of trained models. Our models achieve pre-training perplexity and downstream task performance comparable to traditional masked protein language models, demonstrating that interpretability does not compromise performance. While adaptable to any language model, we focus on masked protein language models due to their importance in drug discovery and the ability to validate our model's capabilities through real-world experiments and expert knowledge. We scale our CB-pLM from 24 million to 3 billion parameters, making them the largest Concept Bottleneck Models trained and the first capable of generative language modeling.

2678. Discovering Clone Negatives via Adaptive Contrastive Learning for Image-Text Matching

链接：https://iclr.cc/virtual/2025/poster/29908 abstract： In this paper, we identify a common yet challenging issue in image-text matching, i.e., clone negatives: negative image-text pairs that semantically resemble positive pairs, leading to ambiguous and sub-optimal matching outcomes. To tackle this issue, we propose Adaptive Contrastive Learning (AdaCL), which introduces two margin parameters along with a modulating anchor to dynamically strengthen the compactness between positives and mitigate the influence of clone negatives. The modulating anchor is selected based on the distribution of negative samples without the need for explicit training, allowing for progressive tuning and advanced in-batch supervision. Extensive experiments across several tasks demonstrate the effectiveness of AdaCL in image-text matching. Furthermore, we extend AdaCL to weakly-supervised image-text matching by replacing human-annotated descriptions with automatically generated captions, thereby increasing the number of potential clone negatives. AdaCL maintains robustness in this setting, alleviating the reliance on crowd-sourced annotations and laying a foundation for scalable vision-language contrastive learning.

2679. PostCast: Generalizable Postprocessing for Precipitation Nowcasting via Unsupervised Blurriness Modeling

链接：https://iclr.cc/virtual/2025/poster/27936 abstract： Precipitation nowcasting plays a pivotal role in socioeconomic sectors, especially in severe convective weather warnings. Although notable progress has been achieved by approaches mining the spatiotemporal correlations with deep learning, these methods still suffer severe blurriness as the lead time increases, which hampers accurate predictions for extreme precipitation. To alleviate blurriness, researchers explore generative methods conditioned on blurry predictions. However, the pairs of blurry predictions and corresponding ground truth need to be given in advance, making the training pipeline cumbersome and limiting the generality of generative models within blurry modes that appear in training data. By rethinking the blurriness in precipitation nowcasting as a blur kernel acting on predictions, we propose an unsupervised postprocessing method to eliminate the blurriness without the requirement of training with the pairs of blurry predictions and corresponding ground truth. Specifically, we utilize blurry predictions to guide the generation process of a pre-trained unconditional denoising diffusion probabilistic model (DDPM) to obtain high-fidelity predictions with eliminated blurriness. A zero-shot blur kernel estimation mechanism and an auto-scale denoise guidance strategy are introduced to adapt the unconditional DDPM to any blurriness modes varying from datasets and lead times in precipitation nowcasting. Extensive experiments are conducted on 7 precipitation radar datasets, demonstrating the generality and superiority of our method.

2680. Mitigating Reward Over-Optimization in RLHF via Behavior-Supported Regularization

链接：https://iclr.cc/virtual/2025/poster/29757 abstract： Reinforcement learning from human feedback (RLHF) is an effective method for aligning large language models (LLMs) with human values. However, reward over-optimization remains an open challenge leading to discrepancies between the performance of LLMs under the reward model and the true human objectives. A primary contributor to reward over-optimization is the extrapolation error that arises when the reward model evaluates out-of-distribution (OOD) responses. However, current methods still fail to prevent the increasing frequency of OOD response generation during the reinforcement learning (RL) process and are not effective at handling extrapolation errors from OOD responses. In this work, we propose the Behavior-Supported Policy Optimization (BSPO) method to mitigate the reward over-optimization issue. Specifically, we define behavior policy as the next token distribution of the reward training dataset to model the in-distribution (ID) region of the reward model. Building on this, we introduce the behavior-supported Bellman operator to regularize the value function, penalizing all OOD values without impacting the ID ones. Consequently, BSPO reduces the generation of OOD responses during the RL process, thereby avoiding overestimation caused by the reward model’s extrapolation errors. Theoretically, we prove that BSPO guarantees a monotonic improvement of the supported policy until convergence to the optimal behavior-supported policy. Empirical results from extensive experiments show that BSPO outperforms baselines in preventing reward over-optimization due to OOD evaluation and finding the optimal ID policy.

2681. A Closer Look at Machine Unlearning for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29714 abstract： Large language models (LLMs) may memorize sensitive or

copyrighted content, raising privacy and legal concerns. Due to the high cost of retraining from scratch, researchers attempt to employ machine unlearning to remove specific content from LLMs while preserving the overall performance. In this paper, we discuss several issues in machine unlearning for LLMs and provide our insights on possible approaches. To address the issue of inadequate evaluation of model outputs after unlearning, we introduce three additional metrics to evaluate token diversity, sentence semantics, and factual correctness. We then categorize unlearning methods into untargeted and targeted, and discuss their issues respectively. Specifically, the behavior that untargeted unlearning attempts to approximate is unpredictable and may involve hallucinations, and existing regularization is insufficient for targeted unlearning. To alleviate these issues, we propose using the objective of maximizing entropy (ME) for untargeted unlearning and incorporate answer preservation (AP) loss as regularization for targeted unlearning. Experimental results across three scenarios, i.e., fictitious unlearning, continual unlearning, and real-world unlearning, demonstrate the effectiveness of our approaches. The code is available at https://github.com/sail-sg/closer-look-LLM-unlearning.

2682. Looking Inward: Language Models Can Learn About Themselves by Introspection

链接：https://iclr.cc/virtual/2025/poster/28917 abstract： Humans acquire knowledge by observing the external world, but also by introspection. Introspection gives a person privileged access to their current state of mind (e.g. thoughts and feelings) that are not accessible to external observers. Do LLMs have this introspective capability of privileged access? If they do, this would show that LLMs can acquire knowledge not contained in or inferable from training data.We investigate LLMs predicting properties of their own behavior in hypothetical situations. If a model M1 has this capability, it should outperform a different model M2 in predicting M1's behavior—even if M2 is trained on M1's ground-truth behavior.The idea is that M1 has privileged access to its own behavioral tendencies, and this enables it to predict itself better than M2 (even if M2 is generally stronger).In experiments with GPT-4, GPT-4o, and Llama-3 models, we find that the model M1 outperforms M2 in predicting itself, providing evidence for privileged access. Further experiments and ablations provide additional evidence.Our results show that LLMs can offer reliable self-information independent of external data in certain domains. By demonstrating this, we pave the way for further work on introspection in more practical domains, which would have significant implications for model transparency and explainability. However, while we successfully show introspective capabilities in simple tasks, we are unsuccessful on more complex tasks or those requiring out-of-distribution generalization.

2683. Perplexity Trap: PLM-Based Retrievers Overrate Low Perplexity Documents

链接：https://iclr.cc/virtual/2025/poster/29494 abstract： Previous studies have found that PLM-based retrieval models exhibit a preference for LLM-generated content, assigning higher relevance scores to these documents even when their semantic quality is comparable to human-written ones. This phenomenon, known as source bias, threatens the sustainable development of the information access ecosystem. However, the underlying causes of source bias remain unexplored. In this paper, we explain the process of information retrieval with a causal graph and discover that PLM-based retrievers learn perplexity features for relevance estimation, causing source bias by ranking the documents with low perplexity higher. Theoretical analysis further reveals that the phenomenon stems from the positive correlation between the gradients of the loss functions in language modeling task and retrieval task. Based on the analysis, a causal-inspired inference-time debiasing method is proposed, called Causal Diagnosis and Correction (CDC). CDC first diagnoses the bias effect of the perplexity and then separates the bias effect from the overall estimated relevance score. Experimental results across three domains demonstrate the superior debiasing effectiveness of CDC, emphasizing the validity of our proposed explanatory framework. Source codes are available at https://github.com/WhyDwelledOnAi/Perplexity-Trap.

2684. Influence-Guided Diffusion for Dataset Distillation

链接：https://iclr.cc/virtual/2025/poster/31231 abstract： Dataset distillation aims to streamline the training process by creating a compact yet effective dataset for a much larger original dataset. However, existing methods often struggle with distilling large, high-resolution datasets due to prohibitive resource costs and limited performance, primarily stemming from sample-wise optimizations in the pixel space. Motivated by the remarkable capabilities of diffusion generative models in learning target dataset distributions and controllably sampling high-quality data tailored to user needs, we propose framing dataset distillation as a controlled diffusion generation task aimed at generating data specifically tailored for effective training purposes. By establishing a correlation between the overarching objective of dataset distillation and the trajectory influence function, we introduce the Influence-Guided Diffusion (IGD) sampling framework to generate training-effective data without the need to retrain diffusion models. An efficient guided function is designed by leveraging the trajectory influence function as an indicator to steer diffusions to produce data with influence promotion and diversity enhancement. Extensive experiments show that the training performance of distilled datasets generated by diffusions can be significantly improved by integrating with our IGD method and achieving state-of-the-art performance in distilling ImageNet datasets. Particularly, an exceptional result is achieved on the ImageNet-1K, reaching 60.3\% at IPC=50. Our code is available at https://github.com/mchen725/DD_IGD.

2685. IterComp: Iterative Composition-Aware Feedback Learning from Model Gallery for Text-to-Image Generation

链接：https://iclr.cc/virtual/2025/poster/30983 abstract： Advanced diffusion models like Stable Diffusion 3, Omost, and FLUX have made notable strides in compositional text-to-image generation. However, these methods typically exhibit distinct strengths for compositional generation, with some excelling in handling attribute binding and others in spatial relationships. This disparity highlights the need for an approach that can leverage the complementary strengths of various models to comprehensively improve the composition capability. To this end, we introduce IterComp, a novel framework that aggregates composition-aware model preferences from multiple models and employs an iterative feedback learning approach to enhance compositional generation. Specifically, we curate a gallery of six powerful open-source diffusion models and evaluate their three key compositional metrics: attribute binding, spatial relationships, and non-spatial relationships. Based on these metrics, we develop a composition-aware model preference dataset comprising numerous image-rank pairs to train composition-aware reward models. Then, we propose an iterative feedback learning method to enhance compositionality in a closed-loop manner, enabling the progressive self-refinement of both the base diffusion model and reward models over multiple iterations. Detailed theoretical proof demonstrates the effectiveness of this method. Extensive experiments demonstrate our significant superiority over previous methods, particularly in multi-category object composition and complex semantic alignment. IterComp opens new research avenues in reward feedback learning for diffusion models and compositional generation. Code: https://github.com/YangLing0818/IterComp

2686. MaxCutPool: differentiable feature-aware Maxcut for pooling in graph neural networks

链接：https://iclr.cc/virtual/2025/poster/27765 abstract： We propose a novel approach to compute the MAXCUT in attributed graphs, i.e., graphs with features associated with nodes and edges. Our approach works well on any kind of graph topology and can find solutions that jointly optimize the MAXCUT along with other objectives. Based on the obtained MAXCUT partition, we implement a hierarchical graph pooling layer for Graph Neural Networks, which is sparse, trainable end-to-end, and particularly suitable for downstream tasks on heterophilic graphs.

2687. DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming Heads

链接：https://iclr.cc/virtual/2025/poster/29062 abstract： Deploying long-context large language models (LLMs) is essential but poses significant computational and memory challenges.Caching all Key and Value (KV) states across all attention heads consumes substantial memory.Existing KV cache pruning methods either damage the long-context capabilities of LLMs or offer only limited efficiency improvements.In this paper, we identify that only a fraction of attention heads, a.k.a, Retrieval Heads, are critical for processing long contexts and require full attention across all tokens.In contrast, all other heads, which primarily focus on recent tokens and attention sinks—referred to as Streaming Heads—do not require full attention.Based on this insight, we introduce DuoAttention, a framework that only applies a full KV cache to retrieval heads while using a light-weight, constant-length KV cache for streaming heads, which reduces both LLM's decoding and pre-filling memory and latency without compromising its long-context abilities.DuoAttention uses a lightweight, optimization-based algorithm with synthetic data to identify retrieval heads accurately.Our method significantly reduces long-context inference memory by up to 2.55$\times$ for MHA and 1.67$\times$ for GQA models while speeding up decoding by up to 2.18$\times$ and 1.50$\times$ and accelerating pre-filling by up to 1.73$\times$ and 1.63$\times$ for MHA and GQA models, respectively, with minimal accuracy loss compared to full attention.Notably, combined with quantization, DuoAttention enables Llama-3-8B decoding with 3.33 million context length measured on a single A100 GPU. Code is provided in https://github.com/mit-han-lab/duo-attention.

2688. Glad: A Streaming Scene Generator for Autonomous Driving

链接：https://iclr.cc/virtual/2025/poster/29229 abstract： The generation and simulation of diverse real-world scenes have significant application value in the field of autonomous driving, especially for the corner cases. Recently, researchers have explored employing neural radiance fields or diffusion models to generate novel views or synthetic data under driving scenes. However, these approaches suffer from unseen scenes or restricted video length, thus lacking sufficient adaptability for data generation and simulation. To address these issues, we propose a simple yet effective framework, named Glad, to generate video data in a frame-by-frame style. To ensure the temporal consistency of synthetic video, we introduce a latent variable propagation module, which views the latent features of previous frame as noise prior and injects it into the latent features of current frame. In addition, we design a streaming data sampler to orderly sample the original image in a video clip at continuous iterations. Given the reference frame, our Glad can be viewed as a streaming simulator by generating the videos for specific scenes. Extensive experiments are performed on the widely-used nuScenes dataset. Experimental results demonstrate that our proposed Glad achieves promising performance, serving as a strong baseline for online video generation. We will release the source code and models publicly.

2689. DreamBench++: A Human-Aligned Benchmark for Personalized Image Generation

链接：https://iclr.cc/virtual/2025/poster/31023 abstract： Personalized image generation holds great promise in assisting humans in everyday work and life due to its impressive function in creatively generating personalized content. However, current evaluations either are automated but misalign with humans or require human evaluations that are time-consuming and

expensive. In this work, we present DreamBench++, a human-aligned benchmark that advanced multimodal GPT models automate. Specifically, we systematically design the prompts to let GPT be both human-aligned and self-aligned, empowered with task reinforcement. Further, we construct a comprehensive dataset comprising diverse images and prompts. By benchmarking 7 modern generative models, we demonstrate that \dreambench results in significantly more human-aligned evaluation, helping boost the community with innovative findings.

2690. Resolution Attack: Exploiting Image Compression to Deceive Deep Neural Networks

链接：https://iclr.cc/virtual/2025/poster/32090 abstract： Model robustness is essential for ensuring the stability and reliability of machine learning systems. Despite extensive research on various aspects of model robustness, such as adversarial robustness and label noise robustness, the exploration of robustness towards different resolutions, remains less explored. To address this gap, we introduce a novel form of attack: the resolution attack. This attack aims to deceive both classifiers and human observers by generating images that exhibit different semantics across different resolutions. To implement the resolution attack, we propose an automated framework capable of generating dual-semantic images in a zero-shot manner. Specifically, we leverage large-scale diffusion models for their comprehensive ability to construct images and propose a staged denoising strategy to achieve a smoother transition across resolutions. Through the proposed framework, we conduct resolution attacks against various off-the-shelf classifiers. The experimental results exhibit high attack success rate, which not only validates the effectiveness of our proposed framework but also reveals the vulnerability of current classifiers towards different resolutions. Additionally, our framework, which incorporates features from two distinct objects, serves as a competitive tool for applications such as face swapping and facial camouflage. The code is available at https://github.com/ywj1/resolution-attack.

2691. Stem-OB: Generalizable Visual Imitation Learning with Stem-Like Convergent Observation through Diffusion Inversion

链接：https://iclr.cc/virtual/2025/poster/27776 abstract： Visual imitation learning methods demonstrate strong performance, yet they lack generalization when faced with visual input perturbations like variations in lighting and textures. This limitation hampers their practical application in real-world settings. To address this, we propose Stem-OB that leverages the inversion process of pretrained image diffusion models to suppress low-level visual differences while maintaining high-level scene structures. This image inversion process is akin to transforming the observation into a shared representation, from which other observations also stem. Stem-OB offers a simple yet effective plug-and-play solution that stands in contrast to data augmentation approaches. It demonstrates robustness to various unspecified appearance changes without the need for additional training. We provide theoretical insights and empirical results that validate the efficacy of our approach in simulated and real settings. Stem-OB shows an exceptionally significant improvement in real-world robotic tasks, where challenging light and appearance changes are present, with an average increase of 22.2% in success rates compared to the best baseline. Please refer to this link for more videos and details.

2692. Circuit Transformer: A Transformer That Preserves Logical Equivalence

链接：https://iclr.cc/virtual/2025/poster/28560 abstract： Implementing Boolean functions with circuits consisting of logic gates is fundamental in digital computer design. However, the implemented circuit must be exactly equivalent, which hinders generative neural approaches on this task due to their occasionally wrong predictions. In this study, we introduce a generative neural model, the “Circuit Transformer”, which eliminates such wrong predictions and produces logic circuits strictly equivalent to given Boolean functions. The main idea is a carefully designed decoding mechanism that builds a circuit step-by-step by generating tokens, which has beneficial “cutoff properties” that block a candidate token once it invalidate equivalence. In such a way, the proposed model works similar to typical LLMs while logical equivalence is strictly preserved. A Markov decision process formulation is also proposed for optimizing certain objectives of circuits. Experimentally, we trained an 88-million-parameter Circuit Transformer to generate equivalent yet more compact forms of input circuits, outperforming existing neural approaches on both synthetic and real world benchmarks, without any violation of equivalence constraints.Code: https://github.com/snowkylin/circuit-transformer

2693. Information Theoretic Text-to-Image Alignment

链接：https://iclr.cc/virtual/2025/poster/29462 abstract： Diffusion models for Text-to-Image (T2I) conditional generation have recently achievedtremendous success. Yet, aligning these models with user’s intentions still involves alaborious trial-and-error process, and this challenging alignment problem has attractedconsiderable attention from the research community. In this work, instead of relying onfine-grained linguistic analyses of prompts, human annotation, or auxiliary vision-languagemodels, we use Mutual Information (MI) to guide model alignment. In brief, our methoduses self-supervised fine-tuning and relies on a point-wise MI estimation between promptsand images to create a synthetic fine-tuning set for improving model alignment. Ouranalysis indicates that our method is superior to the state-of-the-art, yet it only requiresthe pre-trained denoising network of the T2I model itself to estimate MI, and a simplefine-tuning strategy that improves alignment while maintaining image quality. Code available at https://github.com/Chao0511/mitune.

2694. GeoILP: A Synthetic Dataset to Guide Large-Scale Rule Induction

链接：https://iclr.cc/virtual/2025/poster/29037 abstract： Inductive logic programming (ILP) is a machine learning approach aiming to learn explanatory rules from data. While existing ILP systems can successfully solve small-scale tasks, large-scale applications with various language biases are rarely explored. Besides, it is crucial for a large majority of current ILP systems to require expert-defined language bias, which hampers the development of ILP towards broader utilizations. In this paper, we introduce GeoILP, a large-scale synthetic dataset of diverse ILP tasks involving numerous aspects of language bias. These tasks are built from geometry problems, at the level from textbook exercise to regional International Mathematical Olympiad (IMO), with the help of a deduction engine. These problems are elaborately selected to cover all challenging language biases, such as recursion, predicate invention, and high arity. Experimental results show that no existing method can solve GeoILP tasks. In addition, along with classic symbolic-form data, we provide image-form data to boost the development of the joint learning of neural perception and symbolic rule induction.

2695. Rethinking Reward Model Evaluation: Are We Barking up the Wrong Tree?

链接：https://iclr.cc/virtual/2025/poster/30496 abstract： Reward Models (RMs) are crucial for aligning language models with human preferences. Currently, the evaluation of RMs depends on measuring accuracy against a validation set of manually annotated preference data.Although this method is straightforward and widely adopted, the relationship between RM accuracy and downstream policy performance remains under-explored.In this work, we conduct experiments in a synthetic setting to investigate how differences in RM measured by accuracy translate into gaps in optimized policy performance.Our findings reveal that while there is a weak positive correlation between accuracy and downstream performance, policies optimized towards RMs with similar accuracy can exhibit quite different performance.Moreover, we discover that the way of measuring accuracy significantly impacts its ability to predict the final policy performance. Through the lens of the Regressional Goodhart effect, we recognize that accuracy, when used for measuring RM quality, can fail to fully capture the potential RM overoptimization.This underscores the inadequacy of relying solely on accuracy to reflect their impact on policy optimization.

2696. CREIMBO: Cross-Regional Ensemble Interactions in Multi-view Brain Observations

链接：https://iclr.cc/virtual/2025/poster/31159 abstract： Modern recordings of neural activity provide diverse observations of neurons across brain areas, behavioral conditions, and subjects; presenting an exciting opportunity to reveal the fundamentals of brain-wide dynamics. Current analysis methods, however, often fail to fully harness the richness of such data, as they provide either uninterpretable representations (e.g., via deep networks) or oversimplify models (e.g., by assuming stationary dynamics or analyzing each session independently). Here, instead of regarding asynchronous neural recordings that lack alignment in neural identity or brain areas as a limitation, we leverage these diverse views into the brain to learn a unified model of neural dynamics. Specifically, we assume that brain activity is driven by multiple hidden global sub-circuits. These sub-circuits represent global basis interactions between neural ensembles—functional groups of neurons—such that the time-varying decomposition of these sub-circuits defines how the ensembles' interactions evolve over time non-stationarily and non-linearly.We discover the neural ensembles underlying non-simultaneous observations, along with their non-stationary evolving interactions, with our new model, CREIMBO (Cross-Regional Ensemble Interactions in Multi-view Brain Observations). CREIMBO identifies the hidden composition of per-session neural ensembles through novel graph-driven dictionary learning and models the ensemble dynamics on a low-dimensional manifold spanned by a sparse time-varying composition of the global sub-circuits. Thus, CREIMBO disentangles overlapping temporal neural processes while preserving interpretability due to the use of a shared underlying sub-circuit basis. Moreover, CREIMBO distinguishes session-specific computations from global (session-invariant) ones by identifying session covariates and variations in sub-circuit activations. We demonstrate CREIMBO's ability to recover true components in synthetic data, and uncover meaningful brain dynamics in human high-density electrode recordings, including cross-subject neural mechanisms as well as inter- vs. intra-region dynamical motifs. Furthermore, using mouse whole-brain recordings, we show CREIMBO's ability to discover dynamical interactions that capture task and behavioral variables and meaningfully align with the biological importance of the brain areas they represent.

2697. EvA: Erasing Spurious Correlations with Activations

链接：https://iclr.cc/virtual/2025/poster/27672 abstract： Spurious correlations often arise when models associate features strongly correlated with, but not causally related to, the label e.g. an image classifier associates bodies of water with ducks. To mitigate spurious correlations, existing methods focus on learning unbiased representation or incorporating additional information about the correlations during training. This work removes spurious correlations by ``Erasing with Activations'' (EvA). EvA learns class-specific spurious indicator on each channel for the fully connected layer of pretrained networks. By erasing spurious connections during re-weighting, EvA achieves state-of-the-art performance across diverse datasets (6.2\% relative gain on BAR and achieves 4.1\% on Waterbirds). For biased datasets without any information about the spurious correlations, EvA can outperform previous methods (4.8\% relative gain on Waterbirds) with 6 orders of magnitude less compute, highlighting its data and computational efficiency.

2698. SysBench: Can LLMs Follow System Message?

链接：https://iclr.cc/virtual/2025/poster/30048 abstract： Large Language Models (LLMs) have become instrumental across various applications, with the customization of these models to specific scenarios becoming increasingly critical. System message, a fundamental component of LLMs, is consist of carefully crafted instructions that guide the behavior of model to meet intended goals. Despite the recognized potential of system messages to optimize AI-driven solutions, there is a notable absence of a comprehensive benchmark for evaluating how well LLMs follow system messages. To fill this gap, we introduce SysBench, a benchmark that systematically analyzes system message following ability in terms of three limitations of existing LLMs: constraint violation, instruction misjudgement and multi-turn instability. Specifically, we manually construct evaluation dataset based on six prevalent types of constraints, including 500 tailor-designed system messages and multi-turn user conversations covering various interaction relationships. Additionally, we develop a comprehensive evaluation protocol to measure model performance. Finally, we conduct extensive evaluation across various existing LLMs, measuring their ability to follow specified constraints given in system messages. The results highlight both the strengths and weaknesses of existing models, offering key insights and directions for future research.

2699. Walk the Talk? Measuring the Faithfulness of Large Language Model Explanations

链接：https://iclr.cc/virtual/2025/poster/30987 abstract： Large language models (LLMs) are capable of generating plausible explanations of how they arrived at an answer to a question. However, these explanations can misrepresent the model's "reasoning" process, i.e., they can be unfaithful. This, in turn, can lead to over-trust and misuse. We introduce a new approach for measuring the faithfulness of LLM explanations. First, we provide a rigorous definition of faithfulness. Since LLM explanations mimic human explanations, they often reference high-level concepts in the input question that purportedly influenced the model. We define faithfulness in terms of the difference between the set of concepts that the LLM's explanations imply are influential and the set that truly are. Second, we present a novel method for estimating faithfulness that is based on: (1) using an auxiliary LLM to modify the values of concepts within model inputs to create realistic counterfactuals, and (2) using a hierarchical Bayesian model to quantify the causal effects of concepts at both the example- and dataset-level. Our experiments show that our method can be used to quantify and discover interpretable patterns of unfaithfulness. On a social bias task, we uncover cases where LLM explanations hide the influence of social bias. On a medical question answering task, we uncover cases where LLM explanations provide misleading claims about which pieces of evidence influenced the model's decisions.

2700. Federated Class-Incremental Learning: A Hybrid Approach Using Latent Exemplars and Data-Free Techniques to Address Local and Global Forgetting

链接：https://iclr.cc/virtual/2025/poster/27710 abstract： Federated Class-Incremental Learning (FCIL) refers to a scenario where a dynamically changing number of clients collaboratively learn an ever-increasing number of incoming tasks. FCIL is known to suffer from local forgetting due to class imbalance at each client and global forgetting due to class imbalance across clients. We develop a mathematical framework for FCIL that formulates local and global forgetting. Then, we propose an approach called Hybrid Rehearsal (HR), which utilizes latent exemplars and data-free techniques to address local and global forgetting, respectively. HR employs a customized autoencoder designed for both data classification and the generation of synthetic data. To determine the embeddings of new tasks for all clients in the latent space of the encoder, the server uses the Lennard-Jones Potential formulations. Meanwhile, at the clients, the decoder decodes the stored low-dimensional latent space exemplars back to the high-dimensional input space, used to address local forgetting. To overcome global forgetting, the decoder generates synthetic data. Furthermore, our mathematical framework proves that our proposed approach HR can, in principle, tackle the two local and global forgetting challenges. In practice, extensive experiments demonstrate that while preserving privacy, our proposed approach outperforms the state-of-the-art baselines on multiple FCIL benchmarks with low compute and memory footprints.

2701. ProtPainter: Draw or Drag Protein via Topology-guided Diffusion

链接：https://iclr.cc/virtual/2025/poster/29857 abstract： Recent advances in protein backbone generation have achieved promising results under structural, functional, or physical constraints. However, existing methods lack the flexibility for precise topology control, limiting navigation of the backbone space. We present $\textbf{ProtPainter}$, a diffusion-based approach for generating protein backbones conditioned on 3D curves. ProtPainter follows a two-stage process: curve-based sketching and sketch-guided backbone generation. For the first stage, we propose $\textbf{CurveEncoder}$, which predicts secondary structure annotations from a curve to parametrize sketch generation. For the second stage, the sketch guides the generative process in Denoising Diffusion Probabilistic Modeling (DDPM) to generate backbones. During the process, we further introduce a fusion scheduling scheme, Helix-Gating, to control the scaling factors. To evaluate, we propose the first benchmark for topology-conditioned protein generation, introducing Protein Restoration Task and a new metric, self-consistency Topology Fitness (scTF). Experiments demonstrate ProtPainter's ability to generate topology-fit (scTF $>$ 0.8) and designable (scTM $>$ 0.5) backbones, with drawing and dragging tasks showcasing its flexibility and versatility.

2702. Redefining the task of Bioactivity Prediction

链接：https://iclr.cc/virtual/2025/poster/29625 abstract： Small molecules are vital to modern medicine, and accurately

predicting their bioactivity against protein targets is crucial for therapeutic discovery and development. However, current machine learning models often rely on spurious features, leading to biased outcomes. Notably, a simple pocket-only baseline can achieve results comparable to, and sometimes better than, more complex models that incorporate both the protein pockets and the small molecules. Our analysis reveals that this phenomenon arises from insufficient training data and an improper evaluation process, which is typically conducted at the pocket level rather than the small molecule level. To address these issues, we redefine the bioactivity prediction task by introducing the SIU dataset-a million-scale Structural small molecule-protein Interaction dataset for Unbiased bioactivity prediction task, which is 50 times larger than the widely used PDBbind. The bioactivity labels in SIU are derived from wet experiments and organized by label types, ensuring greater accuracy and comparability. The complexes in SIU are constructed using a majority vote from three commonly used docking software programs, enhancing their reliability. Additionally, the structure of SIU allows for multiple small molecules to be associated with each protein pocket, enabling the redefinition of evaluation metrics like Pearson and Spearman correlations across different small molecules targeting the same protein pocket. Experimental results demonstrate that this new task provides a more challenging and meaningful benchmark for training and evaluating bioactivity prediction models, ultimately offering a more robust assessment of model performance.

2703. CtD: Composition through Decomposition in Emergent Communication

链接：https://iclr.cc/virtual/2025/poster/30034 abstract： Compositionality is a cognitive mechanism that allows humans to systematically combine known concepts in novel ways.This study demonstrates how artificial neural agents acquire and utilize compositional generalization to describe previously unseen images.Our method, termed `Composition through Decomposition'', involves two sequential training steps.In the \'Decompose\' step, the agents learn to decompose an image into basic concepts using a codebook acquired during interaction in a multi-target coordination game.Subsequently, in theCompose\' step, the agents employ this codebook to describe novel images by composing basic concepts into complex phrases.Remarkably, we observe cases where generalization in the `Compose' step is achieved zero-shot, without the need for additional training.

2704. Reframing Structure-Based Drug Design Model Evaluation via Metrics Correlated to Practical Needs

链接：https://iclr.cc/virtual/2025/poster/29635 abstract： Recent advances in structure-based drug design (SBDD) have produced surprising results, with models often generating molecules that achieve better Vina docking scores than actual ligands. However, these results are frequently overly optimistic due to the limitations of docking score accuracy and the challenges of wet-lab validation. While generated molecules may demonstrate high QED (drug-likeness) and SA (synthetic accessibility) scores, they often lack true drug-like properties or synthesizability. To address these limitations, we propose a model-level evaluation framework that emphasizes practical metrics aligned with real-world applications. Inspired by recent findings on the utility of generated molecules in ligand-based virtual screening, our framework evaluates SBDD models by their ability to produce molecules that effectively retrieve active compounds from chemical libraries via similarity-based searches. This approach provides a direct indication of therapeutic potential, bridging the gap between theoretical performance and real-world utility. Our experiments reveal that while SBDD models may excel in theoretical metrics like Vina scores, they often fall short in these practical metrics. By introducing this new evaluation strategy, we aim to enhance the relevance and impact of SBDD models for pharmaceutical research and development.

2705. Rethinking Diffusion Posterior Sampling: From Conditional Score Estimator to Maximizing a Posterior

链接：https://iclr.cc/virtual/2025/poster/30272 abstract： Recent advancements in diffusion models have been leveraged to address inverse problems without additional training, and Diffusion Posterior Sampling (DPS) (Chung et al., 2022a) is among the most popular approaches. Previous analyses suggest that DPS accomplishes posterior sampling by approximating the conditional score. While in this paper, we demonstrate that the conditional score approximation employed by DPS is not as effective as previously assumed, but rather aligns more closely with the principle of maximizing a posterior (MAP). This assertion is substantiated through an examination of DPS on 512$\times$512 ImageNet images, revealing that: 1) DPS's conditional score estimation significantly diverges from the score of a well-trained conditional diffusion model and is even inferior to the unconditional score; 2) The mean of DPS's conditional score estimation deviates significantly from zero, rendering it an invalid score estimation; 3) DPS generates high-quality samples with significantly lower diversity. In light of the above findings, we posit that DPS more closely resembles MAP than a conditional score estimator, and accordingly propose the following enhancements to DPS: 1) we explicitly maximize the posterior through multi-step gradient ascent and projection; 2) we utilize a light-weighted conditional score estimator trained with only 100 images and 8 GPU hours. Extensive experimental results indicate that these proposed improvements significantly enhance DPS's performance. The source code for these improvements is provided in https://github.com/tongdaxu/Rethinking-Diffusion-Posterior-Sampling-From-Conditional-Score-Estimator-to-Maximizing-a-Posterior.

2706. Mask in the Mirror: Implicit Sparsification

链接：https://iclr.cc/virtual/2025/poster/29491 abstract： Continuous sparsification strategies are among the most effective methods for reducing the inference costs and memory demands of large-scale neural networks. A key factor in their success is the implicit L_1 regularization induced by jointly learning both mask and weight variables, which has been shown experimentally to outperform explicit L_1 regularization. We provide a theoretical explanation for this observation by analyzing the learning dynamics, revealing that early continuous sparsification is governed by an implicit L_2 regularization that gradually transitions to an L_1 penalty over time. Leveraging this insight, we propose a method to dynamically control the strength of this implicit bias. Through an extension of the mirror flow framework, we establish convergence and optimality guarantees in the context of underdetermined linear regression. Our theoretical findings may be of independent interest, as we demonstrate how to enter the rich regime and show that the implicit bias can be controlled via a time-dependent Bregman potential. To validate these insights, we introduce PILoT, a continuous sparsification approach with novel initialization and dynamic regularization, which consistently outperforms baselines in standard experiments.

2707. The Rise and Down of Babel Tower: Investigating the Evolution Process of Multilingual Code Large Language Model

链接：https://iclr.cc/virtual/2025/poster/28901 abstract： Large language models (LLMs) have shown significant multilingual capabilities. However, the mechanisms underlying the development of these capabilities during pre-training are not well understood. In this paper, we use code LLMs as an experimental platform to explore the evolution of multilingual capabilities in LLMs during the pre-training process. Based on our observations, we propose the Babel Tower Hypothesis, which describes the entire process of LLMs acquiring new language capabilities. During the learning process, multiple languages initially share a single knowledge system dominated by the primary language and gradually develop language-specific knowledge systems. We then validate the above hypothesis by tracking the internal states of the LLM using specific methods. Experimental results show that the internal state changes of the LLM are consistent with our Babel Tower Hypothesis. Building on these insights, we propose a novel method to construct an optimized pre-training corpus for multilingual code LLMs, which significantly outperforms LLMs trained on the original corpus. The proposed Babel Tower Hypothesis provides new insights into designing pre-training data distributions to achieve optimal multilingual capabilities in LLMs.

2708. Score-based Self-supervised MRI Denoising

链接：https://iclr.cc/virtual/2025/poster/27982 abstract：

2709. Revisit Micro-batch Clipping: Adaptive Data Pruning via Gradient Manipulation

链接：https://iclr.cc/virtual/2025/poster/28314 abstract： Micro-batch clipping, a gradient clipping method, has recently shown potential in enhancing auto-speech recognition (ASR) model performance. However, the underlying mechanism behind this improvement remains mysterious, particularly the observation that only certain micro-batch sizes are beneficial. In this paper, we make the first attempt to explain this phenomenon. Inspired by recent data pruning research, we assume that specific training samples may impede model convergence during certain training phases. Under this assumption, the convergence analysis shows that micro-batch clipping can improve the convergence rate asymptotically at the cost of an additional constant bias that does not diminish with more training iterations. The bias is dependent on a few factors and can be minimized at specific micro-batch size, thereby elucidating the existence of the sweet-spot micro-batch size observed previously. We also verify the effectiveness of micro-batch clipping beyond speech models on vision and language models, and show promising performance gains in these domains. An exploration of potential limitations shows that micro-batch clipping is less effective when training data originates from multiple distinct domains.

2710. Contextualizing biological perturbation experiments through language

链接：https://iclr.cc/virtual/2025/poster/30942 abstract： High-content perturbation experiments allow scientists to probe biomolecular systems at unprecedented resolution, but experimental and analysis costs pose significant barriers to widespread adoption. Machine learning has the potential to guide efficient exploration of the perturbation space and extract novel insights from these data. However, current approaches neglect the semantic richness of the relevant biology, and their objectives are misaligned with downstream biological analyses. In this paper, we hypothesize that large language models (LLMs) present a natural medium for representing complex biological relationships and rationalizing experimental outcomes. We propose PerturbQA, a benchmark for structured reasoning over perturbation experiments. Unlike current benchmarks that primarily interrogate existing knowledge, PerturbQA is inspired by open problems in perturbation modeling: prediction of differential expression and change of direction for unseen perturbations, and gene set enrichment. We evaluate state-of-the-art machine learning and statistical approaches for modeling perturbations, as well as standard LLM reasoning strategies, and we find that current methods perform poorly on PerturbQA. As a proof of feasibility, we introduce Summer (SUMMarize, retrievE, and answeR, a simple, domain-informed LLM framework that matches or exceeds the current state-of-the-art. Our code and data are publicly available at https://github.com/genentech/PerturbQA.

2711. PhyMPGN: Physics-encoded Message Passing Graph Network for spatiotemporal PDE systems

链接：https://iclr.cc/virtual/2025/poster/28878 abstract： Solving partial differential equations (PDEs) serves as a cornerstone for modeling complex dynamical systems. Recent progresses have demonstrated grand benefits of data-driven neural-based models for predicting spatiotemporal dynamics (e.g., tremendous speedup gain compared with classical numerical methods). However, most existing neural models rely on rich training data, have limited extrapolation and generalization abilities, and suffer to produce precise or reliable physical prediction under intricate conditions (e.g., irregular mesh or geometry, complex boundary conditions, diverse PDE parameters, etc.). To this end, we propose a new graph learning approach, namely, Physics-encoded Message Passing Graph Network (PhyMPGN), to model spatiotemporal PDE systems on irregular meshes given small training datasets. Specifically, we incorporate a GNN into a numerical integrator to approximate the temporal marching of spatiotemporal dynamics for a given PDE system. Considering that many physical phenomena are governed by diffusion processes, we further design a learnable Laplace block, which encodes the discrete Laplace-Beltrami operator, to aid and guide the GNN learning in a physically feasible solution space. A boundary condition padding strategy is also designed to improve the model convergence and accuracy. Extensive experiments demonstrate that PhyMPGN is capable of accurately predicting various types of spatiotemporal dynamics on coarse unstructured meshes, consistently achieves the state-of-the-art results, and outperforms other baselines with considerable gains.

2712. Local Patterns Generalize Better for Novel Anomalies

链接：https://iclr.cc/virtual/2025/poster/30988 abstract： Video anomaly detection (VAD) aims to identify novel actions or events which are unseen during training. Existing mainstream VAD techniques typically focus on the global patterns with redundant details and struggle to generalize to unseen samples. In this paper, we propose a framework that identifies the local patterns which generalize to novel samples and models the dynamics of local patterns. The capability of extracting spatial local patterns is achieved through a two-stage process involving image-text alignment and cross-modality attention. Generalizable representations are built by focusing on semantically relevant components which can be recombined to capture the essence of novel anomalies, reducing unnecessary visual data variances. To enhance local patterns with temporal clues, we propose a State Machine Module (SMM) that utilizes earlier high-resolution textual tokens to guide the generation of precise captions for subsequent low-resolution observations. Furthermore, temporal motion estimation complements spatial local patterns to detect anomalies characterized by novel spatial distributions or distinctive dynamics. Extensive experiments on popular benchmark datasets demonstrate the achievement of state-of-the-art performance. Code is available at https://github.com/AllenYLJiang/Local-Patterns-Generalize-Better/.

2713. MS-Diffusion: Multi-subject Zero-shot Image Personalization with Layout Guidance

链接：https://iclr.cc/virtual/2025/poster/29759 abstract： Recent advancements in text-to-image generation models have dramatically enhanced the generation of photorealistic images from textual prompts, leading to an increased interest in personalized text-to-image applications, particularly in multi-subject scenarios. However, these advances are hindered by two main challenges: firstly, the need to accurately maintain the details of each referenced subject in accordance with the textual descriptions; and secondly, the difficulty in achieving a cohesive representation of multiple subjects in a single image without introducing inconsistencies. To address these concerns, our research introduces the MS-Diffusion framework for layout-guided zero-shot image personalization with multi-subjects. This innovative approach integrates grounding tokens with the feature resampler to maintain detail fidelity among subjects. With the layout guidance, MS-Diffusion further improves the cross-attention to adapt to the multi-subject inputs, ensuring that each subject condition acts on specific areas. The proposed multi-subject cross-attention orchestrates harmonious inter-subject compositions while preserving the control of texts. Comprehensive quantitative and qualitative experiments affirm that this method surpasses existing models in both image and text fidelity, promoting the development of personalized text-to-image generation.

2714. Model Risk-sensitive Offline Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28781 abstract： Offline reinforcement learning (RL) is becoming critical in risk-sensitive areas such as finance and autonomous driving, where incorrect decisions can lead to substantial financial loss or compromised safety. However, traditional risk-sensitive offline RL methods often struggle with accurately assessing risk, with minor errors in the estimated return potentially causing significant inaccuracies of risk estimation. These challenges are intensified by distribution shifts inherent in offline RL. To mitigate these issues, we propose a model risk-sensitive offline RL framework designed to minimize the worst-case of risks across a set of plausible alternative scenarios rather than solely focusing on minimizing estimated risk. We present a critic-ensemble criterion method that identifies the plausible alternative scenarios without introducing additional hyperparameters. We also incorporate the learned Fourier feature framework and the IQN framework to address spectral bias in neural networks, which can otherwise lead to severe errors in calculating model risk. Our experiments in finance and self-driving scenarios demonstrate that the proposed framework significantly reduces risk, by 11.2% to 18.5%, compared to the most outperforming risk-sensitive offline RL baseline, particularly in highly uncertain environments.

2715. Beyond Squared Error: Exploring Loss Design for Enhanced Training of Generative Flow Networks

链接：https://iclr.cc/virtual/2025/poster/31016 abstract： Generative Flow Networks (GFlowNets) are a novel class of

generative models designed to sample from unnormalized distributions and have found applications in various important tasks, attracting great research interest in their training algorithms. In general, GFlowNets are trained by fitting the forward flow to the backward flow on sampled training objects. Prior work focused on the choice of training objects, parameterizations, sampling and resampling strategies, and backward policies, aiming to enhance credit assignment, exploration, or exploitation of the training process. However, the choice of regression loss, which can highly influence the exploration and exploitation behavior of the under-training policy, has been overlooked. Due to the lack of theoretical understanding for choosing an appropriate regression loss, most existing algorithms train the flow network by minimizing the squared error of the forward and backward flows in log-space, i.e., using the quadratic regression loss. In this work, we rigorously prove that distinct regression losses correspond to specific divergence measures, enabling us to design and analyze regression losses according to the desired properties of the corresponding divergence measures. Specifically, we examine two key properties: zero-forcing and zero-avoiding, where the former promotes exploitation and higher rewards, and the latter encourages exploration and enhances diversity. Based on our theoretical framework, we propose three novel regression losses, namely, Shifted-Cosh, Linex(1/2), and Linex(1). We evaluate them across three benchmarks: hyper-grid, bit-sequence generation, and molecule generation. Our proposed losses are compatible with most existing training algorithms, and significantly improve the performances of the algorithms concerning convergence speed, sample diversity, and robustness.

2716. Offline RL in Regular Decision Processes: Sample Efficiency via Language Metrics

链接：https://iclr.cc/virtual/2025/poster/30400 abstract： This work studies offline Reinforcement Learning (RL) in a class of non-Markovian environments called Regular Decision Processes (RDPs). In RDPs, the unknown dependency of future observations and rewards from the past interactions can be captured by some hidden finite-state automaton. For this reason, many RDP algorithms first reconstruct this unknown dependency using automata learning techniques. In this paper, we consider episodic RDPs and show that it is possible to overcome the limitations of existing offline RL algorithms for RDPs viathe introduction of two original techniques: a novel metric grounded in formal language theory and an approach based on Count-Min-Sketch (CMS). Owing to the novel language metric, our algorithm is proven to be more sample efficient than existing results, and in some problem instances admitting low complexity languages, the gain is showcased to be exponential in the episode length. The CMS-based approach removes the need for naïve counting and alleviates the memory requirements for long planning horizons. We derive Probably Approximately Correct (PAC) sample complexity bounds associated to each of these techniques, and validate the approach experimentally.

2717. TempMe: Video Temporal Token Merging for Efficient Text-Video Retrieval

链接：https://iclr.cc/virtual/2025/poster/28516 abstract： Most text-video retrieval methods utilize the text-image pre-trained models like CLIP as a backbone. These methods process each sampled frame independently by the image encoder, resulting in high computational overhead and limiting practical deployment. Addressing this, we focus on efficient text-video retrieval by tackling two key challenges: 1. From the perspective of trainable parameters, current parameter-efficient fine-tuning methods incur high inference costs; 2. From the perspective of model complexity, current token compression methods are mainly designed for images to reduce spatial redundancy but overlook temporal redundancy in consecutive frames of a video. To tackle these challenges, we propose Temporal Token Merging (TempMe), a parameter-efficient and training-inference efficient text-video retrieval architecture that minimizes trainable parameters and model complexity. Specifically, we introduce a progressive multi-granularity framework. By gradually combining neighboring clips, we reduce spatio-temporal redundancy and enhance temporal modeling across different frames, leading to improved efficiency and performance. Extensive experiments validate the superiority of our TempMe. Compared to previous parameter-efficient text-video retrieval methods, TempMe achieves superior performance with just 0.50M trainable parameters. It significantly reduces output tokens by 95% and GFLOPs by 51%, while achieving a 1.8X speedup and a 4.4% R-Sum improvement. With full fine-tuning, TempMe achieves a significant 7.9% R-Sum improvement, trains 1.57X faster, and utilizes 75.2% GPU memory usage. The code is available at https://github.com/LunarShen/TempMe.

2718. CViT: Continuous Vision Transformer for Operator Learning

链接：https://iclr.cc/virtual/2025/poster/29048 abstract： Operator learning, which aims to approximate maps between infinite-dimensional function spaces, is an important area in scientific machine learning with applications across various physical domains. Here we introduce the Continuous Vision Transformer (CViT), a novel neural operator architecture that leverages advances in computer vision to address challenges in learning complex physical systems. CViT combines a vision transformer encoder, a novel grid-based coordinate embedding, and a query-wise cross-attention mechanism to effectively capture multi-scale dependencies. This design allows for flexible output representations and consistent evaluation at arbitrary resolutions. We demonstrate CViT's effectiveness across a diverse range of partial differential equation (PDE) systems, including fluid dynamics, climate modeling, and reaction-diffusion processes. Our comprehensive experiments show that CViT achieves state-of-the-art performance on multiple benchmarks, often surpassing larger foundation models, even without extensive pretraining and roll-out fine-tuning. Taken together, CViT exhibits robust handling of discontinuous solutions, multi-scale features, and intricate spatio-temporal dynamics. Our contributions can be viewed as a significant step towards adapting advanced computer vision architectures for building more flexible and accurate machine learning models in the physical sciences.

2719. Rethinking Multiple-Instance Learning From Feature Space to Probability Space

链接：https://iclr.cc/virtual/2025/poster/28025 abstract： Multiple-instance learning (MIL) was initially proposed to identify key instances within a set (bag) of instances when only one bag-level label is provided. Current deep MIL models mostly solve multi-instance problem in feature space. Nevertheless, with the increasing complexity of data, we found this paradigm faces significant risks in representation learning stage, which could lead to algorithm degradation in deep MIL models. We speculate that the degradation issue stems from the persistent drift of instances in feature space during learning. In this paper, we propose a novel Probability-Space MIL network (PSMIL) as a countermeasure. In PSMIL, a self-training alignment strategy is introduced in probability space to cope with the drift problem in feature space, and the alignment target objective is proven mathematically optimal. Furthermore, we reveal that the widely-used attention-based pooling mechanism in current deep MIL models is easily affected by the perturbation in feature space and further introduce an alternative called probability-space attention pooling. It effectively captures the key instance in each bag from feature space to probability space, and further eliminates the impact of selection drift in the pooling stage. To summarize, PSMIL seeks to solve a MIL problem in probability space rather than feature space. Experimental results illustrate that PSMIL could potentially achieve performance close to supervised learning level in complex tasks (gap within 5\%), with the incremental alignment in propability space bring more than 19\% accuracy improvements for current existing mainstream models in simulated CIFAR datasets. For existing publicly available MIL benchmarks/datasets, attention in probability space also achieves competitive performance to the state-of-the-art deep MIL models. Codes are available at \url{https://github.com/LMBDA-design/PSAMIL}.

2720. SVG: 3D Stereoscopic Video Generation via Denoising Frame Matrix

链接：https://iclr.cc/virtual/2025/poster/28086 abstract： Video generation models have demonstrated great capability of producing impressive monocular videos, however, the generation of 3D stereoscopic video remains under-explored. We propose a pose-free and training-free approach for generating 3D stereoscopic videos using an off-the-shelf monocular video generation model. Our method warps a generated monocular video into camera views on stereoscopic baseline using estimated video depth, and employs a novel frame matrix video inpainting framework. The framework leverages the video generation model to inpaint frames observed from different timestamps and views. This effective approach generates consistent and semantically coherent stereoscopic videos without scene optimization or model fine-tuning. Moreover, we develop a disocclusion boundary re-injection scheme that further improves the quality of video inpainting by alleviating the negative effects propagated from disoccluded areas in the latent space. We validate the efficacy of our proposed method by conducting experiments on videos from various generative models, including Sora [4], Lumiere [2], WALT [8], and Zeroscope [12]. The experiments demonstrate that our method has a significant improvement over previous methods. Project page at https://daipengwa.github.io/SVG_ProjectPage/

2721. AgentOccam: A Simple Yet Strong Baseline for LLM-Based Web Agents

链接：https://iclr.cc/virtual/2025/poster/28353 abstract： Autonomy via agents based on large language models (LLMs) that can carry out personalized yet standardized tasks presents a significant opportunity to drive human efficiency. There is an emerging need and interest in automating web tasks (e.g., booking a hotel for a given date within a budget). Being a practical use case itself, the web agent also serves as an important proof-of-concept example for various agent grounding scenarios, with its success promising advancements in many future applications. Meanwhile, much prior research focuses on handcrafting their web agent strategies (e.g., agent's prompting templates, reflective workflow, role-play and multi-agent systems, search or sampling methods, etc.) and the corresponding in-context examples. However, these custom strategies often struggle with generalizability across all potential real-world applications. On the other hand, there has been limited study on the misalignment between a web agent's observation and action representation, and the data on which the agent's underlying LLM has been pre-trained. This discrepancy is especially notable when LLMs are primarily trained for language completion rather than tasks involving embodied navigation actions and symbolic web elements. In our study, we enhance an LLM-based web agent by simply refining its observation and action space, aligning these more closely with the LLM's capabilities. This approach enables our base agent to significantly outperform previous methods on a wide variety of web tasks. Specifically, on WebArena, a benchmark featuring general-purpose web interaction tasks, our agent AgentOccam surpasses the previous state-of-the-art and concurrent work by 9.8 (+29.4%) and 5.9 (+15.8%) absolute points respectively, and boosts the success rate by 26.6 points (+161%) over similar plain web agents with its observation and action space alignment. Furthermore, on WebVoyager benchmark comprising tasks defined on real-world websites, AgentOccam exceeds the former best agent by 2.4 points (+4.6%) on tasks with deterministic answers. We achieve this without using in-context examples, new agent roles, online feedback or search strategies. AgentOccam's simple design highlights LLMs' impressive zero-shot performance on web tasks, and underlines the critical role of carefully tuning observation and action spaces for LLM-based agents.

2722. TabWak: A Watermark for Tabular Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/30853 abstract： Synthetic data offers alternatives for data augmentation and sharing. Till date, it remains unknown how to use watermarking techniques to trace and audit synthetic tables generated by tabular diffusion models to mitigate potential misuses. In this paper, we design TabWak, the first watermarking method to embed invisible signatures that control the sampling of Gaussian latent codes used to synthesize table rows via the diffusion backbone.

TabWak has two key features. Different from existing image watermarking techniques, TabWak uses self-cloning and shuffling to embed the secret key in positional information of random seeds that control the Gaussian latents, allowing to use different seeds at each row for high inter-row diversity and enabling row-wise detectability. To further boost the robustness of watermark detection against post-editing attacks, TabWak uses a valid-bit mechanism that focuses on the tail of the latent code distribution for superior noise resilience. We provide theoretical guarantees on the row diversity and effectiveness of detectability. We evaluate TabWak on five datasets against baselines to show that the quality of watermarked tables remains nearly indistinguishable from non-watermarked tables while achieving high detectability in the presence of strong post-editing attacks, with a 100% true positive rate at a 0.1% false positive rate on synthetic tables with fewer than 300 rows. Our code is available at the following anonymized repository https://github.com/chaoyitud/TabWak.

2723. Interpretable Unsupervised Joint Denoising and Enhancement for Real-World low-light Scenarios

链接：https://iclr.cc/virtual/2025/poster/29750 abstract： Real-world low-light images often suffer from complex degradations such as local overexposure, low brightness, noise, and uneven illumination. Supervised methods tend to overfit to specific scenarios, while unsupervised methods, though better at generalization, struggle to model these degradations due to the lack of reference images. To address this issue, we propose an interpretable, zero-reference joint denoising and low-light enhancement framework tailored for real-world scenarios. Our method derives a training strategy based on paired sub-images with varying illumination and noise levels, grounded in physical imaging principles and retinex theory. Additionally, we leverage the Discrete Cosine Transform (DCT) to perform frequency domain decomposition in the sRGB space, and introduce an implicit-guided hybrid representation strategy that effectively separates intricate compounded degradations. In the backbone network design, we develop retinal decomposition network guided by implicit degradation representation mechanisms. Extensive experiments demonstrate the superiority of our method. Code will be available at https://github.com/huaqlili/unsupervised-light-enhance-ICLR2025.

2724. FIRING-Net: A filtered feature recycling network for speech enhancement

链接：https://iclr.cc/virtual/2025/poster/29549 abstract： Current deep neural networks for speech enhancement (SE) aim to minimize the distance between the output signal and the clean target by filtering out noise features from input features. However, when noise and speech components are highly similar, SE models struggle to learn effective discrimination patterns. To address this challenge, we propose a Filter-Recycle-Interguide framework termed Filter-Recycle-INterGuide NETwork (FIRING-Net) for SE, which filters the input features to extract target features and recycles the filtered-out features as non-target features. These two feature sets then guide each other to refine the features, leading to the aggregation of speech information within the target features and noise information within the non-target features. The proposed FIRING-Net mainly consists of a Local Module (LM) and a Global Module (GM). The LM uses outputs of the speech extraction network as target features and the residual between input and output as non-target features. The GM leverages the energy distribution of self-attention map to extract target and non-target features guided by highest and lowest energy regions. Both LM and GM include interaction modules to leverage the two feature sets in an inter-guided manner for collecting speech from non-target features and filtering out noise from target features. Experiments confirm the effectiveness of the Filter-Recycle-Interguide framework, with FIRING-Net achieving a strong balance between SE performance and computational efficiency, surpassing comparable models across various SNR levels and noise environments.

2725. Can We Talk Models Into Seeing the World Differently?

链接：https://iclr.cc/virtual/2025/poster/28696 abstract： Unlike traditional vision-only models, vision language models (VLMs) offer an intuitive way to access visual content through language prompting by combining a large language model (LLM) with a vision encoder. However, both the LLM and the vision encoder come with their own set of biases, cue preferences, and shortcuts, which have been rigorously studied in uni-modal models. A timely question is how such (potentially misaligned) biases and cue preferences behave under multi-modal fusion in VLMs. As a first step towards a better understanding, we investigate a particularly well-studied vision-only bias - the texture vs. shape bias and the dominance of local over global information. As expected, we find that VLMs inherit this bias to some extent from their vision encoders. Surprisingly, the multi-modality alone proves to have important effects on the model behavior, i.e., the joint training and the language querying change the way visual cues are processed. While this direct impact of language-informed training on a model's visual perception is intriguing, it raises further questions on our ability to actively steer a model's output so that its prediction is based on particular visual cues of the user's choice. Interestingly, VLMs have an inherent tendency to recognize objects based on shape information, which is different from what a plain vision encoder would do. Further active steering towards shape-based classifications through language prompts is however limited. In contrast, active VLM steering towards texture-based decisions through simple natural language prompts is often more successful.

2726. ZooProbe: A Data Engine for Evaluating, Exploring, and Evolving Large-scale Training Data for Multimodal LLMs

链接：https://iclr.cc/virtual/2025/poster/29564 abstract： Multimodal Large Language Models (MLLMs) are thriving through

continuous fine-tuning by LLMs. Driven by the law that "scale is everything", MLLMs expand their training sets during version iterations. In this paper, we propose a large-scale training data engine built around an evaluating-exploring-evolving (E3) loop. Evaluating the data provides insights into its characteristics. Exploring quality rules helps identify which data enhances training. Together, these processes facilitate the systematic evolution of new, high-quality data. With the E3 loop, we introduce ZooProbe, an efficient data engine for MLLMs. First, the problem of data expansion is formalized as a tree of sampling and growth. ZooProbe introduces a small-scale model *zoo* to obtain comprehensive evaluations for child datasets. From multiple perspectives, visual, textual, and multimodal models cover over 50 dimensions of intrinsic and meta attributes, such as object and topic distribution, and higher-level properties, like annotation quality and scene complexity. ZooProbe constructs based on A* search, modeling the heuristic function as a quality estimate from data evaluation results. It dynamically explores the rule of data quality based on the model state of the *probe* datasets. Additionally, it evolves new targeted data with identified high-quality rules. We also develop an extra heuristic quality ranker with the data utilized and discarded during the expansion. Our experiments show that ZooProbe significantly breaks the scaling law in multimodal instruction fine-tuning at scales of 260k and below.ZooProbe generates high-quality data that accelerates MLLM training and enhances performance, automating the evolution of large-scale training data.

2727. DLEFT-MKC: Dynamic Late Fusion Multiple Kernel Clustering with Robust Tensor Learning via Min-Max Optimization

链接：https://iclr.cc/virtual/2025/poster/30238 abstract： Recent advancements in multiple kernel clustering (MKC) have highlighted the effectiveness of late fusion strategies, particularly in enhancing computational efficiency to near-linear complexity while achieving promising clustering performance. However, existing methods encounter three significant limitations: (1) reliance on fixed base partition matrices that do not adaptively optimize during the clustering process, thereby constraining their performance to the inherent representational capabilities of these matrices; (2) a focus on adjusting kernel weights to explore inter-view consistency and complementarity, which often neglects the intrinsic high-order correlations among views, thereby limiting the extraction of comprehensive multiple kernel information; (3) a lack of adaptive mechanisms to accommodate varying distributions within the data, which limits robustness and generalization. To address these challenges, this paper proposes a novel algorithm termed Dynamic Late Fusion Multiple Kernel Clustering with Robust {Tensor Learning via min-max optimization (DLEFT-MKC), which effectively overcomes the representational bottleneck of base partition matrices and facilitates the learning of meaningful high-order cross-view information. Specifically, it is the first to incorporate a min-max optimization paradigm into tensor-based MKC, enhancing algorithm robustness and generalization. Additionally, it dynamically reconstructs decision layers to enhance representation capabilities and subsequently stacks the reconstructed representations for tensor learning that promotes the capture of high-order associations and cluster structures across views, ultimately yielding consensus clustering partitions. To solve the resultant optimization problem, we innovatively design a strategy that combines reduced gradient descent with the alternating direction method of multipliers, ensuring convergence to local optima while maintaining high computational efficiency. Extensive experimental results across various benchmark datasets validate the superior effectiveness and efficiency of the proposed DLEFT-MKC.

2728. Simple yet Effective Incomplete Multi-view Clustering: Similarity-level Imputation and Intra-view Hybrid-group Prototype Construction

链接：https://iclr.cc/virtual/2025/poster/30038 abstract： Most of incomplete multi-view clustering (IMVC) methods typically choose to ignore the missing samples and only utilize observed unpaired samples to construct bipartite similarity. Moreover, they employ a single quantity of prototypes to extract the information of $\textbf{all}$ views. To eliminate these drawbacks, we present a simple yet effective IMVC approach, SIIHPC, in this work. It firstly transforms partial bipartition learning into original sample form by virtue of reconstruction concept to split out of observed similarity, and then loosens traditional non-negative constraints via regularizing samples to more freely characterize the similarity. Subsequently, it learns to recover the incomplete parts by utilizing the connection built between the similarity exclusive on respective view and the consensus graph shared for all views. On this foundation, it further introduces a group of hybrid prototype quantities for each individual view to flexibly extract the data features belonging to each view itself. Accordingly, the resulting graphs are with various scales and describe the overall similarity more comprehensively. It is worth mentioning that these all are optimized in one unified learning framework, which makes it possible for them to reciprocally promote. Then, to effectively solve the formulated optimization problem, we design an ingenious auxiliary function that is with theoretically proven monotonic-increasing properties. Finally, the clustering results are obtained by implementing spectral grouping action on the eigenvectors of stacked multi-scale consensus similarity. Experimental results confirm the effectiveness of SIIHPC.

2729. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality

链接：https://iclr.cc/virtual/2025/poster/31482 abstract： The causal capabilities of large language models (LLMs) are a matter of significant debate, with critical implications for the use of LLMs in societally impactful domains such as medicine, science, law, and policy. We conduct a "behavorial" study of LLMs to benchmark their capability in generating causal arguments. Across a wide range of tasks, we find that LLMs can generate text corresponding to correct causal arguments with high probability, surpassing the best-performing existing methods. Algorithms based on GPT-3.5 and 4 outperform existing algorithms on a pairwise causal discovery task (97%, 13 points gain), counterfactual reasoning task (92%, 20 points gain) and event causality (86% accuracy in determining necessary and sufficient causes in vignettes). We perform robustness checks across tasks and

show that the capabilities cannot be explained by dataset memorization alone, especially since LLMs generalize to novel datasets that were created after the training cutoff date. That said, LLMs exhibit unpredictable failure modes and we discuss the kinds of errors that may be improved and what are the fundamental limits of LLM-based answers. Overall, by operating on the text metadata, LLMs bring capabilities so far understood to be restricted to humans, such as using collected knowledge to generate causal graphs or identifying background causal context from natural language. As a result, LLMs may be used by human domain experts to save effort in setting up a causal analysis, one of the biggest impediments to the widespread adoption of causal methods. Given that LLMs ignore the actual data, our results also point to a fruitful research direction of developing algorithms that combine LLMs with existing causal techniques. Code and datasets are available at https://github.com/py-why/pywhy-llm.

2730. An Online Learning Theory of Trading-Volume Maximization

链接：https://iclr.cc/virtual/2025/poster/29793 abstract： We explore brokerage between traders in an online learning framework.At any round t, two traders meet to exchange an asset, provided the exchange is mutually beneficial.The broker proposes a trading price, and each trader tries to sell their asset or buy the asset from the other party, depending on whether the price is higher or lower than their private valuations.A trade happens if one trader is willing to sell and the other is willing to buy at the proposed price.Previous work provided guidance to a broker aiming at enhancing traders' total earnings by maximizing the *gain from trade*, defined as the sum of the traders' net utilities after each interaction.This classical notion of reward can be highly unfair to traders with small profit margins, and far from the real-life utility of the broker.For these reasons, we investigate how the broker should behave to maximize the trading volume, i.e., the *total number of trades*.We model the traders' valuations as an i.i.d. process with an unknown distribution.If the traders' valuations are revealed after each interaction (full-feedback), and the traders' valuations cumulative distribution function (cdf) is continuous, we provide an algorithm achieving logarithmic regret and show its optimality up to constants.If only their willingness to sell or buy at the proposed price is revealed after each interaction (2-bit feedback), we provide an algorithm achieving poly-logarithmic regret when the traders' valuations cdf is Lipschitz and show its near-optimality.We complement our results by analyzing the implications of dropping the regularity assumptions on the unknown traders' valuations cdf. If we drop the continuous cdf assumption, the regret rate degrades to $\Theta(\sqrt{T})$ in the full-feedback case, where T is the time horizon. If we drop the Lipschitz cdf assumption, learning becomes impossible in the 2-bit feedback case.

2731. UniCBE: An Uniformity-driven Comparing Based Evaluation Framework with Unified Multi-Objective Optimization

链接：https://iclr.cc/virtual/2025/poster/28168 abstract： Human preference plays a significant role in measuring large language models and guiding them to align with human values. Unfortunately, current comparing-based evaluation (CBE) methods typically focus on a single optimization objective, failing to effectively utilize scarce yet valuable preference signals. To address this, we delve into key factors that can enhance the accuracy, convergence, and scalability of CBE: suppressing sampling bias, balancing descending process of uncertainty, and mitigating updating uncertainty.Following the derived guidelines, we propose UniCBE, a unified uniformity-driven CBE framework which simultaneously optimize these core objectives by constructing and integrating three decoupled sampling probability matrices, each designed to ensure uniformity in specific aspects. We further ablate the optimal tuple sampling and preference aggregation strategies to achieve efficient CBE.On the AlpacaEval benchmark, UniCBE saves over 17% of evaluation budgets while achieving a Pearson correlation with ground truth exceeding 0.995, demonstrating excellent accuracy and convergence. In scenarios where new models are continuously introduced, UniCBE can even save over 50% of evaluation costs, highlighting its improved scalability.

2732. Dynamical Diffusion: Learning Temporal Dynamics with Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29071 abstract： Diffusion models have emerged as powerful generative frameworks by progressively adding noise to data through a forward process and then reversing this process to generate realistic samples. While these models have achieved strong performance across various tasks and modalities, their application to temporal predictive learning remains underexplored. Existing approaches treat predictive learning as a conditional generation problem, but often fail to fully exploit the temporal dynamics inherent in the data, leading to challenges in generating temporally coherent sequences. To address this, we introduce Dynamical Diffusion (DyDiff), a theoretically sound framework that incorporates temporally aware forward and reverse processes. Dynamical Diffusion explicitly models temporal transitions at each diffusion step, establishing dependencies on preceding states to better capture temporal dynamics. Through the reparameterization trick, Dynamical Diffusion achieves efficient training and inference similar to any standard diffusion model. Extensive experiments across scientific spatiotemporal forecasting, video prediction, and time series forecasting demonstrate that Dynamical Diffusion consistently improves performance in temporal predictive tasks, filling a crucial gap in existing methodologies. Code is available at this repository: https://github.com/thuml/dynamical-diffusion.

2733. Mitigating Memorization in Language Models

链接：https://iclr.cc/virtual/2025/poster/29943 abstract： Language models (LMs) can “memorize” information, i.e., encode training data in their weights in such a way that inference-time queries can lead to verbatim regurgitation of that data. This ability to extract training data can be problematic, for example, when data are private or sensitive. In this work, we investigate methods

to mitigate memorization: three regularizer-based, three fine-tuning-based, and eleven machine unlearning-based methods, with five of the latter being new methods that we introduce. We also introduce TinyMem, a suite of small, computationally-efficient LMs for the rapid development and evaluation of memorization-mitigation methods. We demonstrate that the mitigation methods that we develop using TinyMem can successfully be applied to production-grade LMs, and we determine via experiment that: regularizer-based mitigation methods are slow and ineffective at curbing memorization; fine-tuning-based methodsare effective at curbing memorization, but overly expensive, especially for retaining higher accuracies; and unlearning-based methods are faster and more effective, allowing for the precise localization and removal of memorized information from LM weights prior to inference. We show, in particular, that our proposed unlearning method BalancedSubnet outperforms other mitigation methods at removingmemorized information while preserving performance on target tasks.

2734. Predicting the Energy Landscape of Stochastic Dynamical System via Physics-informed Self-supervised Learning

链接：https://iclr.cc/virtual/2025/poster/29722 abstract： Energy landscapes play a crucial role in shaping dynamics of many real-world complex systems. System evolution is often modeled as particles moving on a landscape under the combined effect of energy-driven drift and noise-induced diffusion, where the energy governs the long-term motion of the particles.Estimating the energy landscape of a system has been a longstanding interdisciplinary challenge, hindered by the high operational costs or the difficulty of obtaining supervisory signals. Therefore, the question of how to infer the energy landscape in the absence of true energy values is critical. In this paper, we propose a physics-informed self-supervised learning method to learn the energy landscape from the evolution trajectories of the system. It first maps the system state from the observation space to a discrete landscape space by an adaptive codebook, and then explicitly integrates energy into the graph neural Fokker-Planck equation, enabling the joint learning of energy estimation and evolution prediction. Experimental results across interdisciplinary systems demonstrate that our estimated energy has a correlation coefficient above 0.9 with the ground truth, and evolution prediction accuracy exceeds the baseline by an average of 17.65\%. The code is available at https://github.com/tsinghua-fib-lab/PESLA.

2735. Image Watermarks are Removable using Controllable Regeneration from Clean Noise

链接：https://iclr.cc/virtual/2025/poster/32060 abstract： Image watermark techniques provide an effective way to assert ownership, deter misuse, and trace content sources, which has become increasingly essential in the era of large generative models. A critical attribute of watermark techniques is their robustness against various manipulations. In this paper, we introduce a watermark removal approach capable of effectively nullifying state-of-the-art watermarking techniques. Our primary insight involves regenerating the watermarked image starting from a \textbf{clean Gaussian noise} via a controllable diffusion model, utilizing the extracted semantic and spatial features from the watermarked image. The semantic control adapter and the spatial control network are specifically trained to control the denoising process towards ensuring image quality and enhancing consistency between the cleaned image and the original watermarked image. To achieve a smooth trade-off between watermark removal performance and image consistency, we further propose an adjustable and controllable regeneration scheme. This scheme adds varying numbers of noise steps to the latent representation of the watermarked image, followed by a controlled denoising process starting from this noisy latent representation. As the number of noise steps increases, the latent representation progressively approaches clean Gaussian noise, facilitating the desired trade-off. We apply our watermark removal methods across various watermarking techniques, and the results demonstrate that our methods offer superior visual consistency/quality and enhanced watermark removal performance compared to existing regeneration approaches. Our code is available at \url{https://github.com/yepengliu/CtrlRegen}.

2736. Fair Clustering in the Sliding Window Model

链接：https://iclr.cc/virtual/2025/poster/29424 abstract： We study streaming algorithms for proportionally fair clustering, a notion originally suggested by Chierichetti et al. (2017), in the sliding window model. We show that although there exist efficient streaming algorithms in the insertion-only model, surprisingly no algorithm can achieve finite ratio without violating the fairness constraint in sliding window. Hence, the problem of fair clustering is a rare separation between the insertion-only streaming model and the sliding window model. On the other hand, we show that if the fairness constraint is relaxed by a multiplicative $(1+\varepsilon)$ factor, there exists a $(1 + \varepsilon)$-approximate sliding window algorithm that uses $\text{poly}(k\varepsilon^{-1}\log n)$ space. This achieves essentially the best parameters (up to degree in the polynomial) provided the aforementioned lower bound. We also implement a number of empirical evaluations on real datasets to complement our theoretical results.

2737. ELICIT: LLM Augmentation Via External In-context Capability

链接：https://iclr.cc/virtual/2025/poster/30525 abstract： Enhancing the adaptive capabilities of large language models is a critical pursuit in both research and application.Traditional fine-tuning methods require substantial data, computational resources, and specific capabilities, while in-context learning is limited by the need for appropriate demonstrations and efficient token usage. Inspired by the expression of in-context learned capabilities through task vectors and the concept of modular capability or knowledge, we propose ELICIT, a framework consisting of two modules designed to effectively store and reuse task vectors to enhance the diverse adaptive capabilities of models without additional training or inference tokens. Our comprehensive experiments and analysis demonstrate that our pipeline is highly transferable across different input formats,

tasks, and model architectures. Externally storing and reusing vectors that represent in-context learned capabilities not only shows the potential to extract modular capabilities but also significantly enhances the performance, versatility, adaptability, and scalability of large language models, paving the way for more efficient and effective use of these models in a wide range of applications.

2738. Personality Alignment of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/31265 abstract： Aligning large language models (LLMs) typically aim to reflect general human values and behaviors, but they often fail to capture the unique characteristics and preferences of individual users. To address this gap, we introduce the concept of Personality Alignment. This approach tailors LLMs' responses and decisions to match the specific preferences of individual users or closely related groups. Inspired by psychometrics, we created the Personality Alignment with Personality Inventories (PAPI) dataset, which includes data from over 320,000 real subjects across multiple personality assessments - including both the Big Five Personality Factors and Dark Triad traits. This comprehensive dataset enables quantitative evaluation of LLMs' alignment capabilities across both positive and potentially problematic personality dimensions. Recognizing the challenges of personality alignments—such as limited personal data, diverse preferences, and scalability requirements—we developed an activation intervention optimization method. This method enhances LLMs' ability to efficiently align with individual behavioral preferences using minimal data and computational resources. Remarkably, our method, PAS, achieves superior performance while requiring only 1/5 of the optimization time compared to DPO, offering practical value for personality alignment. Our work paves the way for future AI systems to make decisions and reason in truly personality ways, enhancing the relevance and meaning of AI interactions for each user and advancing human-centered artificial intelligence. The dataset and code are released at https://github.com/zhu-minjun/PAlign.

2739. Neural Causal Graph for Interpretable and Intervenable Classification

链接：https://iclr.cc/virtual/2025/poster/28396 abstract： Advancements in neural networks have significantly enhanced the performance of classification models, achieving remarkable accuracy across diverse datasets. However, these models often lack transparency and do not support interactive reasoning with human users, which are essential attributes for applications that require trust and user engagement. To overcome these limitations, we introduce an innovative framework, Neural Causal Graph (NCG), that integrates causal inference with neural networks to enable interpretable and intervenable reasoning. We then propose an intervention training method to model the intervention probability of the prediction, serving as a contextual prompt to facilitate the fine-grained reasoning and human-AI interaction abilities of NCG. Our experiments show that the proposed framework significantly enhances the performance of traditional classification baselines. Furthermore, NCG achieves nearly 95\% top-1 accuracy on the ImageNet dataset by employing a test-time intervention method. This framework not only supports sophisticated post-hoc interpretation but also enables dynamic human-AI interactions, significantly improving the model's transparency and applicability in real-world scenarios.

2740. Language Models are Advanced Anonymizers

链接：https://iclr.cc/virtual/2025/poster/30788 abstract： Recent privacy research on large language models (LLMs) has shown that they achieve near-human-level performance at inferring personal data from online texts. With ever-increasing model capabilities, existing text anonymization methods are currently lacking behind regulatory requirements and adversarial threats. In this work, we take two steps to bridge this gap: First, we present a new setting for evaluating anonymization in the face of adversarial LLM inferences, allowing for a natural measurement of anonymization performance while remedying some of the shortcomings of previous metrics. Then, within this setting, we develop a novel LLM-based adversarial anonymization framework leveraging the strong inferential capabilities of LLMs to inform our anonymization procedure. We conduct a comprehensive experimental evaluation of adversarial anonymization across 13 LLMs on real-world and synthetic online texts, comparing it against multiple baselines and industry-grade anonymizers. Our evaluation shows that adversarial anonymization outperforms current commercial anonymizers both in terms of the resulting utility and privacy. We support our findings with a human study (n=50) highlighting a strong and consistent human preference for LLM-anonymized texts.

2741. Ward: Provable RAG Dataset Inference via LLM Watermarks

链接：https://iclr.cc/virtual/2025/poster/28576 abstract： RAG enables LLMs to easily incorporate external data, raising concerns for data owners regarding unauthorized usage of their content. The challenge of detecting such unauthorized usage remains underexplored, with datasets and methods from adjacent fields being ill-suited for its study. We take several steps to bridge this gap. First, we formalize this problem as (black-box) RAG Dataset Inference (RAG-DI). We then introduce a novel dataset designed for realistic benchmarking of RAG-DI methods, alongside a set of baselines. Finally, we propose Ward, a method for RAG-DI based on LLM watermarks that equips data owners with rigorous statistical guarantees regarding their dataset's misuse in RAG corpora. Ward consistently outperforms all baselines, achieving higher accuracy, superior query efficiency and robustness. Our work provides a foundation for future studies of RAG-DI and highlights LLM watermarks as a promising approach to this problem.

2742. Towards Homogeneous Lexical Tone Decoding from Heterogeneous Intracranial Recordings

链接：https://iclr.cc/virtual/2025/poster/32073 abstract： Recent advancements in brain-computer interfaces (BCIs) and deep learning have made decoding lexical tones from intracranial recordings possible, providing the potential to restore the communication ability of speech-impaired tonal language speakers. However, data heterogeneity induced by both physiological and instrumental factors poses a significant challenge for unified invasive brain tone decoding. Particularly, the existing heterogeneous decoding paradigm (training subject-specific models with individual data) suffers from the intrinsic limitation that fails to learn generalized neural representations and leverages data across subjects. To this end, we introduce Homogeneity-Heterogeneity Disentangled Learning for Neural Representations (H2DiLR), a framework that disentangles and learns the homogeneity and heterogeneity from intracranial recordings of multiple subjects. To verify the effectiveness of H2DiLR, we collected stereoelectroencephalography (sEEG) from multiple participants reading Mandarin materials containing 407 syllables (covering nearly all Mandarin characters). Extensive experiments demonstrate that H2DiLR, as a unified decoding paradigm, outperforms the naive heterogeneous decoding paradigm by a large margin. We also empirically show that H2DiLR indeed captures homogeneity and heterogeneity during neural representation learning.

2743. NutriBench: A Dataset for Evaluating Large Language Models in Nutrition Estimation from Meal Descriptions

链接：https://iclr.cc/virtual/2025/poster/30896 abstract： Accurate nutrition estimation helps people make informed dietary choices and is essential in the prevention of serious health complications. We present NutriBench, the first publicly available natural language meal description nutrition benchmark. NutriBench consists of 11,857 meal descriptions generated from real-world global dietary intake data. The data is human-verified and annotated with macro-nutrient labels, including carbohydrates, proteins, fats, and calories. We conduct an extensive evaluation of Nutribench on the task of carbohydrate estimation, testing twelve leading Large Language Models (LLMs), including GPT-4o, Llama3.1, Qwen2, Gemma2, and OpenBioLLM models, using standard, Chain-of-Thought and Retrieval-Augmented Generation strategies. Additionally, we present a study involving professional nutritionists, finding that LLMs can provide comparable but significantly faster estimates. Finally, we perform a real-world risk assessment by simulating the effect of carbohydrate predictions on the blood glucose levels of individuals with type 1 diabetes. Our work highlights the opportunities and challenges of using LLMs for nutrition estimation, demonstrating their potential to aid professionals and laypersons and improve health outcomes. Our benchmark is publicly available at: https://mehak126.github.io/nutribench.html

2744. SciLitLLM: How to Adapt LLMs for Scientific Literature Understanding

链接：https://iclr.cc/virtual/2025/poster/30757 abstract： Scientific literature understanding is crucial for extracting targeted information and garnering insights, thereby significantly advancing scientific discovery.Despite the remarkable success of Large Language Models (LLMs), they face challenges in scientific literature understanding, primarily due to (1) a lack of scientific knowledge and (2) unfamiliarity with specialized scientific tasks.To develop an LLM specialized in scientific literature understanding, we propose a hybrid strategy that integrates continual pre-training (CPT) and supervised fine-tuning (SFT), to simultaneously infuse scientific domain knowledge and enhance instruction-following capabilities for domain-specific tasks.In this process, we identify two key challenges: (1) constructing high-quality CPT corpora, and (2) generating diverse SFT instructions. We address these challenges through a meticulous pipeline, including PDF text extraction, parsing content error correction, quality filtering, and synthetic instruction creation.Applying this strategy, we present a suite of LLMs: SciLitLLM, specialized in scientific literature understanding.These models demonstrate promising performance on scientific literature understanding benchmarks.(1) We present an effective framework that integrates CPT and SFT to adapt LLMs to scientific literature understanding, which can also be easily adapted to other domains.(2) We propose an LLM-based synthesis method to generate diverse and high-quality scientific instructions, resulting in a new instruction set -- SciLitIns -- for less-represented scientific domains. (3) SciLitLLM achieves promising performance in scientific literature understanding benchmarks.

2745. Linear Spherical Sliced Optimal Transport: A Fast Metric for Comparing Spherical Data

链接：https://iclr.cc/virtual/2025/poster/28865 abstract： Efficient comparison of spherical probability distributions becomes important in fields such as computer vision, geosciences, and medicine. Sliced optimal transport distances, such as spherical and stereographic spherical sliced Wasserstein distances, have recently been developed to address this need. These methods reduce the computational burden of optimal transport by slicing hyperspheres into one-dimensional projections, i.e., lines or circles. Concurrently, linear optimal transport has been proposed to embed distributions into L^2 spaces, where the L^2 distance approximates the optimal transport distance, thereby simplifying comparisons across multiple distributions. In this work, we introduce the Linear Spherical Sliced Optimal Transport (LSSOT) framework, which utilizes slicing to embed spherical distributions into L^2 spaces while preserving their intrinsic geometry, offering a computationally efficient metric for spherical probability measures. We establish the metricity of LSSOT and demonstrate its superior computational efficiency in applications such as cortical surface registration, 3D point cloud interpolation via gradient flow, and shape embedding. Our results demonstrate the significant computational benefits and high accuracy of LSSOT in these applications.

2746. CycleResearcher: Improving Automated Research via Automated Review

链接：https://iclr.cc/virtual/2025/poster/32074 abstract： The automation of scientific discovery has been a long-standing goal within the research community, driven by the potential to accelerate knowledge creation. While significant progress has been made using commercial large language models (LLMs) as research assistants or idea generators, the possibility of automating the entire research process with open-source LLMs remains largely unexplored. This paper explores the feasibility of using open-source post-trained LLMs as autonomous agents capable of performing the full cycle of automated research and review, from literature review and manuscript preparation to peer review and paper refinement. Our iterative preference training framework consists of CycleResearcher, which conducts research tasks, and CycleReviewer, which simulates the peer review process, providing iterative feedback via reinforcement learning. To train these models, we develop two new datasets, Review-5k and Research-14k, reflecting real-world machine learning research and peer review dynamics. Our results demonstrate that CycleReviewer achieves promising performance with a 26.89\% reduction in mean absolute error (MAE) compared to individual human reviewers in predicting paper scores, indicating the potential of LLMs to effectively assist expert-level research evaluation. In research, the papers generated by the CycleResearcher model achieved a score of 5.36 in simulated peer reviews, showing some competitiveness in terms of simulated review scores compared to the preprint level of 5.24 from human experts, while still having room for improvement compared to the accepted paper level of 5.69. This work represents a significant step toward fully automated scientific inquiry, providing ethical safeguards and exploring AI-driven research capabilities. The code, dataset and model weight are released at https://wengsyx.github.io/Researcher.

2747. MMQA: Evaluating LLMs with Multi-Table Multi-Hop Complex Questions

链接：https://iclr.cc/virtual/2025/poster/30290 abstract： While large language models (LLMs) have made strides in understanding tabular data, current tabular evaluation benchmarks, such as WikiTableQuestions and WikiSQL, are focus on single-table scenarios, which cannot necessarily reflect the complexity of real-world applications. To bridge this gap, we present a \textbf{M}ulti-table and Multi-hop Question Answering (MMQA) dataset to assess LLMs' understanding and reasoning capabilities in handling multi-table tasks. The MMQA dataset demands that models perform multiple inferences by drawing evidence from various tables, which are designed to be connected with each other and require models to identify and utilize relationships such as foreign and primary keys. Then, we introduce a comprehensive evaluation framework that tailors to assess LLMs' capabilities in several aspects including Multi-Table Retrieval, Text-to-SQL Generation, Multi-Table QA, Primary Key Selection, and Foreign Key Selection. Finally, we propose a novel multi-table retrieval method that achieves state-of-the-art (SOTA) performance on the MMQA dataset compared to several strong baselines. Our experiment results reveal that, compared with human performance, both open-source and commercial LLMs leave significant performance room for improvements in multi-table understanding and reasoning tasks. We believe that the MMQA benchmark will enhance and facilitate LLMs' multi-table capabilities in real-world scenarios.

2748. Human Simulacra: Benchmarking the Personification of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30585 abstract： Large Language Models (LLMs) are recognized as systems that closely mimic aspects of human intelligence. This capability has attracted the attention of the social science community, who see the potential in leveraging LLMs to replace human participants in experiments, thereby reducing research costs and complexity. In this paper, we introduce a benchmark for LLMs personification, including a strategy for constructing virtual characters' life stories from the ground up, a Multi-Agent Cognitive Mechanism capable of simulating human cognitive processes, and a psychology-guided evaluation method to assess human simulations from both self and observational perspectives. Experimental results demonstrate that our constructed simulacra can produce personified responses that align with their target characters. We hope this work will serve as a benchmark in the field of human simulation, paving the way for future research.

2749. DAMO: Decoding by Accumulating Activations Momentum for Mitigating Hallucinations in Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/30108 abstract： Large Vision-Language Models (VLMs) exhibit significant potential in multimodal tasks but often struggle with hallucinations—responses that are plausible yet visually ungrounded. In this work, we investigate the layer-wise prediction tendencies of VLMs and conduct an in-depth analysis of their decoding mechanism. We observe that VLMs tend to ``overthink'' during the final stages of decoding, making significant prediction shifts in the last few layers often favoring incorrect results, which leads to a surge in hallucinative outputs. Leveraging this localized pattern, we propose a novel decoding strategy inspired by the momentum analogy used in gradient descent-based optimizers. Our method enforces decoding consistency across layers in an adaptive manner during forward passes—an under-explored approach in existing works. This strategy significantly improves the reliability and performance of VLMs in various multimodal tasks, while introducing only negligible efficiency overhead.

2750. UniRestore3D: A Scalable Framework For General Shape Restoration

链接：https://iclr.cc/virtual/2025/poster/27785 abstract： Shape restoration aims to recover intact 3D shapes from defective ones, such as those that are incomplete, noisy, and low-resolution. Previous works have achieved impressive results in shape restoration subtasks thanks to advanced generative models. While effective for specific shape defects, they are less applicable

in real-world scenarios involving multiple defect types simultaneously. Additionally, training on limited subsets of defective shapes hinders knowledge transfer across restoration types and thus affects generalization. In this paper, we address the task of general shape restoration, which restores shapes with various types of defects through a unified model, thereby naturally improving the applicability and scalability. Our approach first standardizes the data representation across different restoration subtasks using high-resolution TSDF grids and constructs a large-scale dataset with diverse types of shape defects. Next, we design an efficient hierarchical shape generation model and a noise-robust defective shape encoder that enables effective impaired shape understanding and intact shape generation. Moreover, we propose a scalable training strategy for efficient model training. The capabilities of our proposed method are demonstrated across multiple shape restoration subtasks and validated on various datasets, including Objaverse, ShapeNet, GSO, and ABO.

2751. Offline Hierarchical Reinforcement Learning via Inverse Optimization

链接：https://iclr.cc/virtual/2025/poster/28982 abstract： Hierarchical policies enable strong performance in many sequential decision-making problems, such as those with high-dimensional action spaces, those requiring long-horizon planning, and settings with sparse rewards. However, learning hierarchical policies from static offline datasets presents a significant challenge.Crucially, actions taken by higher-level policies may not be directly observable within hierarchical controllers, and the offline dataset might have been generated using a different policy structure, hindering the use of standard offline learning algorithms.In this work, we propose $\textit{OHIO}$: a framework for offline reinforcement learning (RL) of hierarchical policies. Our framework leverages knowledge of the policy structure to solve the $\textit{inverse problem}$, recovering the unobservable high-level actions that likely generated the observed data under our hierarchical policy.This approach constructs a dataset suitable for off-the-shelf offline training.We demonstrate our framework on robotic and network optimization problems and show that it substantially outperforms end-to-end RL methods and improves robustness. We investigate a variety of instantiations of our framework, both in direct deployment of policies trained offline and when online fine-tuning is performed. Code and data are available at https://ohio-offline-hierarchical-rl.github.io.

2752. Generative Representational Instruction Tuning

链接：https://iclr.cc/virtual/2025/poster/30586 abstract： All text-based language problems can be reduced to either generation or embedding. Current models only perform well at one or the other. We introduce generative representational instruction tuning (GRIT) whereby a large language model is trained to handle both generative and embedding tasks by distinguishing between them through instructions. Compared to other open models, our resulting GritLM-7B is among the top models on the Massive Text Embedding Benchmark (MTEB) and outperforms various models up to its size on a range of generative tasks. By scaling up further, GritLM-8x7B achieves even stronger generative performance while still being among the best embedding models. Notably, we find that GRIT matches training on only generative or embedding data, thus we can unify both at no performance loss. Among other benefits, the unification via GRIT speeds up Retrieval-Augmented Generation (RAG) by > 60% for long documents, by no longer requiring separate retrieval and generation models. Models, code, etc. are freely available at https://github.com/ContextualAI/gritlm.

2753. Training Nonlinear Transformers for Chain-of-Thought Inference: A Theoretical Generalization Analysis

链接：https://iclr.cc/virtual/2025/poster/28427 abstract： Chain-of-Thought (CoT) is an efficient prompting method that enables the reasoning ability of large language models by augmenting the query using multiple examples with multiple intermediate steps. Despite the empirical success, the theoretical understanding of how to train a Transformer to achieve the CoT ability remains less explored. This is primarily due to the technical challenges involved in analyzing the nonconvex optimization on nonlinear attention models. To the best of our knowledge, this work provides the first theoretical study of training Transformers with nonlinear attention to obtain the CoT generalization capability so that the resulting model can inference on unseen tasks when the input is augmented by examples of the new task. We first quantify the required training samples and iterations to train a Transformer model towards CoT ability. We then prove the success of its CoT generalization on unseen tasks with distribution-shifted testing data. Moreover, we theoretically characterize the conditions for an accurate reasoning output by CoT even when the provided reasoning examples contain noises and are not always accurate. In contrast, in-context learning (ICL), which can be viewed as one-step CoT without intermediate steps, may fail to provide an accurate output when CoT does. These theoretical findings are justified through experiments.

2754. Adversarially Robust Anomaly Detection through Spurious Negative Pair Mitigation

链接：https://iclr.cc/virtual/2025/poster/28075 abstract： Despite significant progress in Anomaly Detection (AD), the robustness of existing detection methods against adversarial attacks remains a challenge, compromising their reliability in critical real-world applications such as autonomous driving. This issue primarily arises from the AD setup, which assumes that training data is limited to a group of unlabeled normal samples, making the detectors vulnerable to adversarial anomaly samples during testing. Additionally, implementing adversarial training as a safeguard encounters difficulties, such as formulating an effective objective function without access to labels. An ideal objective function for adversarial training in AD should promote strong perturbations both within and between the normal and anomaly groups to maximize margin between normal and anomaly distribution. To address these issues, we first propose crafting a pseudo-anomaly group derived from normal group samples.

Then, we demonstrate that adversarial training with contrastive loss could serve as an ideal objective function, as it creates both inter- and intra-group perturbations. However, we notice that spurious negative pairs compromise the conventional contrastive loss for achieving robust AD. Spurious negative pairs are those that should be mapped closely but are erroneously separated. These pairs introduce noise and misguide the direction of inter-group adversarial perturbations. To overcome the effect of spurious negative pairs, we define opposite pairs and adversarially pull them apart to strengthen inter-group perturbations. Experimental results demonstrate our superior performance in both clean and adversarial scenarios, with a 26.1% improvement in robust detection across various challenging benchmark datasets.

2755. An Exploration with Entropy Constrained 3D Gaussians for 2D Video Compression

链接：https://iclr.cc/virtual/2025/poster/30096 abstract： 3D Gaussian Splatting (3DGS) has witnessed its rapid development in novel view synthesis, which attains high quality reconstruction and real-time rendering. At the same time, there is still a gap before implicit neural representation (INR) can become a practical compressor due to the lack of stream decoding and real-time frame reconstruction on consumer-grade hardware. It remains a question whether the fast rendering and partial parameter decoding characteristics of 3DGS are applicable to video compression. To address these challenges, we propose a Toast-like Sliding Window (TSW) orthographic projection for converting any 3D Gaussian model into a video representation model. This method efficiently represents video by leveraging temporal redundancy through a sliding window approach. Additionally, the converted model is inherently stream-decodable and offers a higher rendering frame rate compared to INR methods. Building on TSW, we introduce an end-to-end trainable video compression method, GSVC, which employs deformable Gaussian representation and optical flow guidance to capture dynamic content in videos. Experimental results demonstrate that our method effectively transforms a 3D Gaussian model into a practical video compressor. GSVC further achieves better rate-distortion performance than NeRV on the UVG dataset, while achieving higher frame reconstruction speed (+30%~40% fps) and stream decoding. Code is available at Github

2756. Learning Fine-Grained Representations through Textual Token Disentanglement in Composed Video Retrieval

链接：https://iclr.cc/virtual/2025/poster/27859 abstract： With the explosive growth of video data, finding videos that meet detailed requirements in large datasets has become a challenge. To address this, the composed video retrieval task has been introduced, enabling users to retrieve videos using complex queries that involve both visual and textual information. However, the inherent heterogeneity between the modalities poses significant challenges. Textual data are highly abstract, while video content contains substantial redundancy. The modality gap in information representation makes existing methods struggle with the modality fusion and alignment required for fine-grained composed retrieval. To overcome these challenges, we first introduce FineCVR-1M, a fine-grained composed video retrieval dataset containing 1,010,071 video-text triplets with detailed textual descriptions. This dataset is constructed through an automated process that identifies key concept changes between video pairs to generate textual descriptions for both static and action concepts. For fine-grained retrieval methods, the key challenge lies in understanding the detailed requirements. Text description serves as clear expressions of intent, but it requires models to distinguish subtle differences in the description of video semantics. Therefore, we propose a textual Feature Disentanglement and Cross-modal Alignment framework (FDCA) that disentangles features at both the sentence and token levels. At the sequence level, we separate text features into retained and injected features. At the token level, an Auxiliary Token Disentangling mechanism is proposed to disentangle texts into retained, injected, and excluded tokens. The disentanglement at both levels extracts fine-grained features, which are aligned and fused with the reference video to extract global representations for video retrieval. Experiments on FineCVR-1M dataset demonstrate the superior performance of FDCA. Our code and dataset are available at: https://may2333.github.io/FineCVR/.

2757. Decentralized Sporadic Federated Learning: A Unified Algorithmic Framework with Convergence Guarantees

链接：https://iclr.cc/virtual/2025/poster/29019 abstract： Decentralized federated learning (DFL) captures FL settings where both (i) model updates and (ii) model aggregations are exclusively carried out by the clients without a central server. Existing DFL works have mostly focused on settings where clients conduct a fixed number of local updates between local model exchanges, overlooking heterogeneity and dynamics in communication and computation capabilities. In this work, we propose Decentralized Sporadic Federated Learning ($\texttt{DSpodFL}$), a DFL methodology built on a generalized notion of *sporadicity* in both local gradient and aggregation processes. $\texttt{DSpodFL}$ subsumes many existing decentralized optimization methods under a unified algorithmic framework by modeling the per-iteration (i) occurrence of gradient descent at each client and (ii) exchange of models between client pairs as arbitrary indicator random variables, thus capturing *heterogeneous and time-varying* computation/communication scenarios. We analytically characterize the convergence behavior of $\texttt{DSpodFL}$ for both convex and non-convex models and for both constant and diminishing learning rates, under mild assumptions on the communication graph connectivity, data heterogeneity across clients, and gradient noises. We show how our bounds recover existing results from decentralized gradient descent as special cases. Experiments demonstrate that $\texttt{DSpodFL}$ consistently achieves improved training speeds compared with baselines under various system settings.

2758. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory

链接：https://iclr.cc/virtual/2025/poster/28290 abstract： Recent large language model (LLM)-driven chat assistant systems have integrated memory components to track user-assistant chat histories, enabling more accurate and personalized responses. However, their long-term memory capabilities in sustained interactions remain underexplored. We introduce LongMemEval, a comprehensive benchmark designed to evaluate five core long-term memory abilities of chat assistants: information extraction, multi-session reasoning, temporal reasoning, knowledge updates, and abstention. With 500 meticulously curated questions embedded within freely scalable user-assistant chat histories, LongMemEval presents a significant challenge to existing long-term memory systems, with commercial chat assistants and long-context LLMs showing a 30% accuracy drop on memorizing information across sustained interactions. We then present a unified framework that breaks down the long-term memory design into three stages: indexing, retrieval, and reading. Built upon key experimental insights, we propose several memory design optimizations including session decomposition for value granularity, fact-augmented key expansion for indexing, and time-aware query expansion for refining the search scope. Extensive experiments show that these optimizations greatly improve both memory recall and downstream question answering on LongMemEval. Overall, our study provides valuable resources and guidance for advancing the long-term memory capabilities of LLM-based chat assistants, paving the way toward more personalized and reliable conversational AI. Our benchmark and code are publicly available at https://github.com/xiaowu0162/LongMemEval.

2759. T2V2: A Unified Non-Autoregressive Model for Speech Recognition and Synthesis via Multitask Learning

链接：https://iclr.cc/virtual/2025/poster/29506 abstract： We introduce T2V2 (Text to Voice and Voice to Text), a unified non-autoregressive model capable of performing both automatic speech recognition (ASR) and text-to-speech (TTS) synthesis within the same framework. T2V2 uses a shared Conformer backbone with rotary positional embeddings to efficiently handle these core tasks, with ASR trained using Connectionist Temporal Classification (CTC) loss and TTS using masked language modeling (MLM) loss. The model operates on discrete tokens, where speech tokens are generated by clustering features from a self-supervised learning model. To further enhance performance, we introduce auxiliary tasks: CTC error correction to refine raw ASR outputs using contextual information from speech embeddings, and unconditional speech MLM, enabling classifier free guidance to improve TTS. Our method is self-contained, leveraging intermediate CTC outputs to align text and speech using Monotonic Alignment Search, without relying on external aligners. We perform extensive experimental evaluation to verify the efficacy of the T2V2 framework, achieving state-of-the-art performance on TTS task and competitive performance in discrete ASR.

2760. Moner: Motion Correction in Undersampled Radial MRI with Unsupervised Neural Representation

链接：https://iclr.cc/virtual/2025/poster/29816 abstract： Motion correction (MoCo) in radial MRI is a particularly challenging problem due to the unpredictability of subject movement. Current state-of-the-art (SOTA) MoCo algorithms often rely on extensive high-quality MR images to pre-train neural networks, which constrains the solution space and leads to outstanding image reconstruction results. However, the need for large-scale datasets significantly increases costs and limits model generalization. In this work, we propose Moner, an unsupervised MoCo method that jointly reconstructs artifact-free MR images and estimates accurate motion from undersampled, rigid motion-corrupted k-space data, without requiring any training data. Our core idea is to leverage the continuous prior of implicit neural representation (INR) to constrain this ill-posed inverse problem, facilitating optimal solutions. Specifically, we integrate a quasi-static motion model into the INR, granting its ability to correct subject's motion. To stabilize model optimization, we reformulate radial MRI reconstruction as a back-projection problem using the Fourier-slice theorem. Additionally, we propose a novel coarse-to-fine hash encoding strategy, significantly enhancing MoCo accuracy. Experiments on multiple MRI datasets show our Moner achieves performance comparable to SOTA MoCo techniques on in-domain data, while demonstrating significant improvements on out-of-domain data. The code is available at: https://github.com/iwuqing/Moner

2761. One Hundred Neural Networks and Brains Watching Videos: Lessons from Alignment

链接：https://iclr.cc/virtual/2025/poster/30005 abstract： What can we learn from comparing video models to human brains, arguably the most efficient and effective video processing systems in existence? Our work takes a step towards answering this question by performing the first large-scale benchmarking of deep video models on representational alignment to the human brain, using publicly available models and a recently released video brain imaging (fMRI) dataset. We disentangle four factors of variation in the models (temporal modeling, classification task, architecture, and training dataset) that affect alignment to the brain, which we measure by conducting Representational Similarity Analysis across multiple brain regions and model layers. We show that temporal modeling is key for alignment to brain regions involved in early visual processing, while a relevant classification task is key for alignment to higher-level regions. Moreover, we identify clear differences between the brain scoring patterns across layers of CNNs and Transformers, and reveal how training dataset biases transfer to alignment with functionally selective brain areas. Additionally, we uncover a negative correlation of computational complexity to brain alignment. Measuring

a total of 99 neural networks and 10 human brains watching videos, we aim to forge a path that widens our understanding of temporal and semantic video representations in brains and machines, ideally leading towards more efficient video models and more mechanistic explanations of processing in the human brain.

2762. BrainACTIV: Identifying visuo-semantic properties driving cortical selectivity using diffusion-based image manipulation

链接：https://iclr.cc/virtual/2025/poster/30527 abstract： The human brain efficiently represents visual inputs through specialized neural populations that selectively respond to specific categories. Advancements in generative modeling have enabled data-driven discovery of neural selectivity using brain-optimized image synthesis. However, current methods independently generate one sample at a time, without enforcing structural constraints on the generations; thus, these individual images have no explicit point of comparison, making it hard to discern which image features drive neural response selectivity. To address this issue, we introduce Brain Activation Control Through Image Variation (BrainACTIV), a method for manipulating a reference image to enhance or decrease activity in a target cortical region using pretrained diffusion models. Starting from a reference image allows for fine-grained and reliable offline identification of optimal visuo-semantic properties, as well as producing controlled stimuli for novel neuroimaging studies. We show that our manipulations effectively modulate predicted fMRI responses and agree with hypothesized preferred categories in established regions of interest, while remaining structurally close to the reference image. Moreover, we demonstrate how our method accentuates differences between brain regions that are selective to the same category, and how it could be used to explore neural representation of brain regions with unknown selectivities. Hence, BrainACTIV holds the potential to formulate robust hypotheses about brain representation and to facilitate the production of naturalistic stimuli for neuroscientific experiments.

2763. Multi-Draft Speculative Sampling: Canonical Decomposition and Theoretical Limits

链接：https://iclr.cc/virtual/2025/poster/29904 abstract： We consider multi-draft speculative sampling, where the proposal sequences are sampled independently from different draft models. At each step, a token-level draft selection scheme takes a list of valid tokens as input and produces an output token whose distribution matches that of the target model. Previous works have demonstrated that the optimal scheme (which maximizes the probability of accepting one of the input tokens) can be cast as a solution to a linear program. In this work we show that the optimal scheme can be decomposed into a two-step solution: in the first step an importance sampling (IS) type scheme is used to select one intermediate token; in the second step (single-draft) speculative sampling is applied to generate the output token. For the case of two identical draft models we further 1) establish a necessary and sufficient condition on the distributions of the target and draft models for the acceptance probability to equal one and 2) provide an explicit expression for the optimal acceptance probability. Our theoretical analysis also motives a new class of token-level selection schemes based on weighted importance sampling. Our experimental results demonstrate consistent improvements in the achievable block efficiency and token rates over baseline schemes in a number of scenarios.

2764. TD-Paint: Faster Diffusion Inpainting Through Time-Aware Pixel Conditioning

链接：https://iclr.cc/virtual/2025/poster/28905 abstract： Diffusion models have emerged as highly effective techniques for inpainting, however, they remain constrained by slow sampling rates. While recent advances have enhanced generation quality, they have also increased sampling time, thereby limiting scalability in real-world applications. We investigate the generative sampling process of diffusion-based inpainting models and observe that these models make minimal use of the input condition during the initial sampling steps. As a result, the sampling trajectory deviates from the data manifold, requiring complex synchronization mechanisms to realign the generation process. To address this, we propose Time-aware Diffusion Paint (TD-Paint), a novel approach that adapts the diffusion process by modeling variable noise levels at the pixel level. This technique allows the model to efficiently use known pixel values from the start, guiding the generation process toward the target manifold. By embedding this information early in the diffusion process, TD-Paint significantly accelerates sampling without compromising image quality. Unlike conventional diffusion-based inpainting models, which require a dedicated architecture or an expensive generation loop, TD-Paint achieves faster sampling times without architectural modifications. Experimental results across three datasets show that TD-Paint outperforms state-of-the-art diffusion models while maintaining lower complexity.

2765. Modeling Complex System Dynamics with Flow Matching Across Time and Conditions

链接：https://iclr.cc/virtual/2025/poster/28730 abstract： Modeling the dynamics of complex real-world systems from temporal snapshot data is crucial for understanding phenomena such as gene regulation, climate change, and financial market fluctuations. Researchers have recently proposed a few methods based either on the Schroedinger Bridge or Flow Matching to tackle this problem, but these approaches remain limited in their ability to effectively combine data from multiple time points and different experimental settings. This integration is essential in real-world scenarios where observations from certain combinations of time points and experimental conditions are missing, either because of experimental costs or sensory failure. To address this challenge, we propose a novel method named Multi-Marginal Flow Matching (MMFM). MMFM first constructs a flow using smooth spline-based interpolation across time points and conditions and regresses it with a neural network using the

classifier-free guided Flow Matching framework. This framework allows for the sharing of contextual information about the dynamics across multiple trajectories. We demonstrate the effectiveness of our method on both synthetic and real-world datasets, including a recent single-cell genomics data set with around a hundred chemical perturbations across time points. Our results show that MMFM significantly outperforms existing methods at imputing data at missing time points.

2766. One for all and all for one: Efficient computation of partial Wasserstein distances on the line

链接：https://iclr.cc/virtual/2025/poster/28547 abstract： Partial Wasserstein helps overcoming some of the limitations of Optimal Transport when the distributions at stake differ in mass, contain noise or outliers or exhibit mass mismatches across distribution modes.We introduce PAWL, a novel algorithm designed to efficiently compute exact PArtial Wasserstein distances on the Line. PAWL not only solves the partial transportation problem for a specified amount of mass to be transported, but *for all* admissible mass amounts. This flexibility is valuable for machine learning tasks where the level of noise is uncertain and needs to be determined through cross-validation, for example. By achieving $O(n \log n)$ time complexity for the partial 1-Wasserstein problem on the line, it enables practical applications with large scale datasets. Additionally, we introduce a novel slicing strategy tailored to Partial Wasserstein, which does not permit transporting mass between outliers or noisy data points. We demonstrate the advantages of PAWL in terms of computational efficiency and performance in downstream tasks, outperforming existing (sliced) Partial Optimal Transport techniques.

2767. Learning Dynamics of Deep Matrix Factorization Beyond the Edge of Stability

链接：https://iclr.cc/virtual/2025/poster/30132 abstract： Deep neural networks trained using gradient descent with a fixed learning rate η often operate in the regime of ``edge of stability'' (EOS), where the largest eigenvalue of the Hessian equilibrates about the stability threshold $2/\eta$. In this work, we present a fine-grained analysis of the learning dynamics of (deep) linear networks (DLNs) within the deep matrix factorization loss beyond EOS. For DLNs, loss oscillations beyond EOS follow a period-doubling route to chaos. We theoretically analyze the regime of the 2-period orbit and show that the loss oscillations occur within a small subspace, with the dimension of the subspace precisely characterized by the learning rate. The crux of our analysis lies in showing that the symmetry-induced conservation law for gradient flow, defined as the balancing gap among the singular values across layers, breaks at EOS and decays monotonically to zero. Overall, our results contribute to explaining two key phenomena in deep networks: (i) shallow models and simple tasks do not always exhibit EOS; and (ii) oscillations occur within top features}. We present experiments to support our theory, along with examples demonstrating how these phenomena occur in nonlinear networks and how they differ from those which have benign landscape such as in DLNs.

2768. Beyond Canonicalization: How Tensorial Messages Improve Equivariant Message Passing

链接：https://iclr.cc/virtual/2025/poster/27926 abstract： In numerous applications of geometric deep learning, the studied systems exhibit spatial symmetries and it is desirable to enforce these. For the symmetry of global rotations and reflections, this means that the model should be equivariant with respect to the transformations that form the group of $\mathrm O(d)$.While many approaches for equivariant message passing require specialized architectures, including non-standard normalization layers or non-linearities, we here present a framework based on local reference frames ("local canonicalization") which can be integrated with any architecture without restrictions.We enhance equivariant message passing based on local canonicalization by introducing tensorial messages to communicate geometric information consistently between different local coordinate frames.Our framework applies to message passing on geometric data in Euclidean spaces of arbitrary dimension.We explicitly show how our approach can be adapted to make a popular existing point cloud architecture equivariant. We demonstrate the superiority of tensorial messages and achieve state-of-the-art results on normal vector regression and competitive results on other standard 3D point cloud tasks.

2769. Manifold Induced Biases for Zero-shot and Few-shot Detection of Generated Images

链接：https://iclr.cc/virtual/2025/poster/30810 abstract： Distinguishing between real and AI-generated images, commonly referred to as 'image detection', presents a timely and significant challenge. Despite extensive research in the (semi-)supervised regime, zero-shot and few-shot solutions have only recently emerged as promising alternatives. Their main advantage is in alleviating the ongoing data maintenance, which quickly becomes outdated due to advances in generative technologies. We identify two main gaps: (1) a lack of theoretical grounding for the methods, and (2) significant room for performance improvements in zero-shot and few-shot regimes. Our approach is founded on understanding and quantifying the biases inherent in generated content, where we use these quantities as criteria for characterizing generated images. Specifically, we explore the biases of the implicit probability manifold, captured by a pre-trained diffusion model. Through score-function analysis, we approximate the curvature, gradient, and bias towards points on the probability manifold, establishing criteria for detection in the zero-shot regime. We further extend our contribution to the few-shot setting by employing a mixture-of-experts methodology. Empirical results across 20 generative models demonstrate that our method outperforms current approaches in both zero-shot and few-shot settings. This work advances the theoretical understanding and practical usage of

generated content biases through the lens of manifold analysis.

2770. Steering Masked Discrete Diffusion Models via Discrete Denoising Posterior Prediction

链接：https://iclr.cc/virtual/2025/poster/29801 abstract： Generative modeling of discrete data underlies important applications spanning text-based agents like ChatGPT to the design of the very building blocks of life in protein sequences. However, application domains need to exert control over the generated data by steering the generative process—typically via RLHF—to satisfy a specified property, reward, or affinity metric. In this paper, we study the problem of steering Masked Diffusion Models (MDMs), a recent class of discrete diffusion models that offer a compelling alternative to traditional autoregressive models. We introduce Discrete Denoising Posterior Prediction (DDPP), a novel framework that casts the task of steering pretrained MDMs as a problem of probabilistic inference by learning to sample from a target Bayesian posterior. Our DDPP framework leads to a family of three novel objectives that are all simulation-free, and thus scalable while applying to general non-differentiable reward functions. Empirically, we instantiate DDPP by steering MDMs to perform class-conditional pixel-level image modeling, RLHF-based alignment of MDMs using text based rewards, and finetuning protein language models to generate more diverse secondary structures and shorter proteins. We substantiate our designs via wet-lab validation, where we observe transient expression of reward-optimized protein sequences.

2771. Latent Space Chain-of-Embedding Enables Output-free LLM Self-Evaluation

链接：https://iclr.cc/virtual/2025/poster/28606 abstract： LLM self-evaluation relies on the LLM's own ability to estimate response correctness, which can greatly improve its deployment reliability. In this research track, we propose the Chain-of-Embedding (CoE) in the latent space to enable LLMs to perform output-free self-evaluation. CoE consists of all progressive hidden states produced during the inference time, which can be treated as the latent thinking path of LLMs. We find that when LLMs respond correctly and incorrectly, their CoE features differ, these discrepancies assist us in estimating LLM response correctness. Experiments in four diverse domains and seven LLMs fully demonstrate the effectiveness of our method. Meanwhile, its label-free design intent without any training and millisecond-level computational cost ensure real-time feedback in large-scale scenarios.More importantly, we provide interesting insights into LLM response correctness from the perspective of hidden state changes inside LLMs.

2772. DriveTransformer: Unified Transformer for Scalable End-to-End Autonomous Driving

链接：https://iclr.cc/virtual/2025/poster/29956 abstract： End-to-end autonomous driving (E2E-AD) has emerged as a trend in the field of autonomous driving, promising a data-driven, scalable approach to system design. However, existing E2E-AD methods usually adopt the sequential paradigm of perception-prediction-planning, which leads to cumulative errors and training instability. The manual ordering of tasks also limits the system’s ability to leverage synergies between tasks (for example, planning-aware perception and game-theoretic interactive prediction and planning). Moreover, the dense BEV representation adopted by existing methods brings computational challenges for long-range perception and long-term temporal fusion. To address these challenges, we present DriveTransformer, a simplified E2E-AD framework for the ease of scaling up, characterized by three key features: Task Parallelism (All agent, map, and planning queries direct interact with each other at each block), Sparse Representation (Task queries direct interact with raw sensor features), and Streaming Processing (Task queries are stored and passed as history information). As a result, the new framework is composed of three unified operations: task self-attention, sensor cross-attention, temporal cross-attention, which significantly reduces the complexity of system and leads to better training stability. DriveTransformer achieves state-of-the-art performance in both simulated closed-loop benchmark Bench2Drive and real world open-loop benchmark nuScenes with high FPS.

2773. An Engorgio Prompt Makes Large Language Model Babble on

链接：https://iclr.cc/virtual/2025/poster/28485 abstract： Auto-regressive large language models (LLMs) have yielded impressive performance in many real-world tasks. However, the new paradigm of these LLMs also exposes novel threats. In this paper, we explore their vulnerability to inference cost attacks, where a malicious user crafts Engorgio prompts to intentionally increase the computation cost and latency of the inference process. We design Engorgio, a novel methodology, to efficiently generate adversarial Engorgio prompts to affect the target LLM's service availability. Engorgio has the following two technical contributions. (1) We employ a parameterized distribution to track LLMs' prediction trajectory. (2) Targeting the auto-regressive nature of LLMs' inference process, we propose novel loss functions to stably suppress the appearance of the token, whose occurrence will interrupt the LLM's generation process. We conduct extensive experiments on 13 open-sourced LLMs with parameters ranging from 125M to 30B. The results show that Engorgio prompts can successfully induce LLMs to generate abnormally long outputs (i.e., roughly 2-13$\times$ longer to reach 90\%+ of the output length limit)in a white-box scenario and our real-world experiment demonstrates Engergio's threat to LLM service with limited computing resources.The code is released at https://github.com/jianshuod/Engorgio-prompt.

2774. MMed-RAG: Versatile Multimodal RAG System for Medical Vision

Language Models

链接：https://iclr.cc/virtual/2025/poster/28145 abstract： Artificial Intelligence (AI) has demonstrated significant potential in healthcare, particularly in disease diagnosis and treatment planning. Recent progress in Medical Large Vision-Language Models (Med-LVLMs) has opened up new possibilities for interactive diagnostic tools. However, these models often suffer from factual hallucination, which can lead to incorrect diagnoses. Fine-tuning and retrieval-augmented generation (RAG) have emerged as methods to address these issues. However, the amount of high-quality data and distribution shifts between training data and deployment data limit the application of fine-tuning methods. Although RAG is lightweight and effective, existing RAG-based approaches are not sufficiently general to different medical domains and can potentially cause misalignment issues, both between modalities and between the model and the ground truth. In this paper, we propose a versatile multimodal RAG system, MMed-RAG, designed to enhance the factuality of Med-LVLMs. Our approach introduces a domain-aware retrieval mechanism, an adaptive retrieved contexts selection, and a provable RAG-based preference fine-tuning strategy. These innovations make the RAG process sufficiently general and reliable, significantly improving alignment when introducing retrieved contexts. Experimental results across five medical datasets (involving radiology, ophthalmology, pathology) on medical VQA and report generation demonstrate that MMed-RAG can achieve an average improvement of 43.8% in factual accuracy in the factual accuracy of Med-LVLMs.

2775. SVDQuant: Absorbing Outliers by Low-Rank Component for 4-Bit Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/27906 abstract： Diffusion models can effectively generate high-quality images. However, as they scale, rising memory demands and higher latency pose substantial deployment challenges. In this work, we aim to accelerate diffusion models by quantizing their weights and activations to 4 bits. At such an aggressive level, both weights and activations are highly sensitive, where existing post-training quantization methods like smoothing become insufficient. To overcome this limitation, we propose *SVDQuant*, a new 4-bit quantization paradigm. Different from smoothing, which redistributes outliers between weights and activations, our approach *absorbs* these outliers using a low-rank branch. We first consolidate the outliers by shifting them from activations to weights. Then, we use a high-precision, low-rank branch to take in the weight outliers with Singular Value Decomposition (SVD), while a low-bit quantized branch handles the residuals. This process eases the quantization on both sides. However, naively running the low-rank branch independently incurs significant overhead due to extra data movement of activations, negating the quantization speedup. To address this, we co-design an inference engine *Nunchaku* that fuses the kernels of the low-rank branch into those of the low-bit branch to cut off redundant memory access. It can also seamlessly support off-the-shelf low-rank adapters (LoRAs) without re-quantization. Extensive experiments on SDXL, PixArt-Σ, and FLUX.1 validate the effectiveness of SVDQuant in preserving image quality. We reduce the memory usage for the 12B FLUX.1 models by 3.5×, achieving 3.0× speedup over the 4-bit weight-only quantization (W4A16) baseline on the 16GB laptop 4090 GPU with INT4 precision. On the latest RTX 5090 desktop with Blackwell architecture, we achieve a 3.1× speedup compared to the W4A16 model using NVFP4 precision. Our quantization library and inference engine are available at https://github.com/mit-han-lab/deepcompressor/ and https://github.com/mit-han-lab/nunchaku/, correspondingly.

2776. Hotspot-Driven Peptide Design via Multi-Fragment Autoregressive Extension

链接：https://iclr.cc/virtual/2025/poster/28614 abstract： Peptides, short chains of amino acids, interact with target proteins, making them a unique class of protein-based therapeutics for treating human diseases. Recently, deep generative models have shown great promise in peptide generation. However, several challenges remain in designing effective peptide binders. First, not all residues contribute equally to peptide-target interactions. Second, the generated peptides must adopt valid geometries due to the constraints of peptide bonds. Third, realistic tasks for peptide drug development are still lacking.To address these challenges, we introduce PepHAR, a hot-spot-driven autoregressive generative model for designing peptides targeting specific proteins. Building on the observation that certain hot spot residues have higher interaction potentials, we first use an energy-based density model to fit and sample these key residues. Next, to ensure proper peptide geometry, we autoregressively extend peptide fragments by estimating dihedral angles between residue frames. Finally, we apply an optimization process to iteratively refine fragment assembly, ensuring correct peptide structures.By combining hot spot sampling with fragment-based extension, our approach enables \textit{de novo} peptide design tailored to a target protein and allows the incorporation of key hot spot residues into peptide scaffolds. Extensive experiments, including peptide design and peptide scaffold generation, demonstrate the strong potential of PepHAR in computational peptide binder design. The source code will be available at https://github.com/Ced3-han/PepHAR.

2777. SANA: Efficient High-Resolution Text-to-Image Synthesis with Linear Diffusion Transformers

链接：https://iclr.cc/virtual/2025/poster/29897 abstract： We introduce Sana, a text-to-image framework that can efficiently generate images up to 4096$\times$4096 resolution. Sana can synthesize high-resolution, high-quality images with strong text-image alignment at a remarkably fast speed, deployable on laptop GPU. Core designs include: (1) Deep compression autoencoder: unlike traditional AEs, which compress images only 8$\times$, we trained an AE that can compress images 32$\times$, effectively reducing the number of latent tokens. (2) Linear DiT: we replace all vanilla attention in DiT with linear

attention, which is more efficient at high resolutions without sacrificing quality. (3) Decoder-only text encoder: we replaced T5 with modern decoder-only small LLM as the text encoder and designed complex human instruction with in-context learning to enhance the image-text alignment. (4) Efficient training and sampling: we propose Flow-DPM-Solver to reduce sampling steps, with efficient caption labeling and selection to accelerate convergence. As a result, Sana-0.6B is very competitive with modern giant diffusion model (e.g. Flux-12B), being 20 times smaller and 100+ times faster in measured throughput. Moreover, Sana-0.6B can be deployed on a 16GB laptop GPU, taking less than 1 second to generate a 1024$\times$1024 resolution image. Sana enables content creation at low cost. Code and model will be publicly released upon publication.

2778. Any-step Dynamics Model Improves Future Predictions for Online and Offline Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/30099 abstract： Model-based methods in reinforcement learning offer a promising approach to enhance data efficiency by facilitating policy exploration within a dynamics model. However, accurately predicting sequential steps in the dynamics model remains a challenge due to the bootstrapping prediction, which attributes the next state to the prediction of the current state. This leads to accumulated errors during model roll-out. In this paper, we propose the Any-step Dynamics Model (ADM) to mitigate the compounding error by reducing bootstrapping prediction to direct prediction. ADM allows for the use of variable-length plans as inputs for predicting future states without frequent bootstrapping. We design two algorithms, ADMPO-ON and ADMPO-OFF, which apply ADM in online and offline model-based frameworks, respectively. In the online setting, ADMPO-ON demonstrates improved sample efficiency compared to previous state-of-the-art methods. In the offline setting, ADMPO-OFF not only demonstrates superior performance compared to recent state-of-the-art offline approaches but also offers better quantification of model uncertainty using only a single ADM.

2779. Dynamic Loss-Based Sample Reweighting for Improved Large Language Model Pretraining

链接：https://iclr.cc/virtual/2025/poster/28818 abstract： Pretraining large language models (LLMs) on vast and heterogeneous datasets is crucial for achieving state-of-the-art performance across diverse downstream tasks. However, current training paradigms treat all samples equally, overlooking the importance or relevance of individual samples throughout the training process. Existing reweighting strategies, which primarily focus on group-level data importance, fail to leverage fine-grained instance-level information and do not adapt dynamically to individual sample importance as training progresses. In this paper, we introduce novel algorithms for dynamic, instance-level data reweighting aimed at improving both the efficiency and effectiveness of LLM pretraining. Our methods adjust the weight of each training sample based on its loss value in an online fashion, allowing the model to dynamically focus on more informative or important samples at the current training stage. In particular, our framework allows us to systematically devise reweighting strategies deprioritizing redundant or uninformative data, which we find tend to work best. Furthermore, we develop a new theoretical framework for analyzing the impact of loss-based reweighting on the convergence of gradient-based optimization, providing the first formal characterization of how these strategies affect convergence bounds. We empirically validate our approach across a spectrum of tasks, from pretraining 7B and 1.4B parameter LLMs to smaller-scale language models and linear regression problems, demonstrating that our loss-based reweighting approach can lead to faster convergence and significantly improved performance.

2780. Semantic Temporal Abstraction via Vision-Language Model Guidance for Efficient Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/27665 abstract： Extracting temporally extended skills can significantly improve the efficiency of reinforcement learning (RL) by breaking down complex decision-making problems with sparse rewards into simpler subtasks and enabling more effective credit assignment. However, existing abstraction methods either discover skills in an unsupervised manner, which often lacks semantic information and leads to erroneous or scattered skill extraction results, or require substantial human intervention. In this work, we propose to leverage the extensive knowledge in pretrained Vision-Language Models (VLMs) to progressively guide the latent space after vector quantization to be more semantically meaningful through relabeling each skill. This approach, termed Vision-language model guided Temporal Abstraction (VanTA), facilitates the discovery of more interpretable and task-relevant temporal segmentations from offline data without the need for extensive manual intervention or heuristics. By leveraging the rich information in VLMs, our method can significantly outperform existing offline RL approaches that depend only on limited training data. From a theory perspective, we demonstrate that stronger internal sequential correlations within each sub-task, induced by VanTA, effectively reduces suboptimality in policy learning. We validate the effectiveness of our approach through extensive experiments on diverse environments, including Franka Kitchen, Minigrid, and Crafter. These experiments show that our method surpasses existing approaches in long-horizon offline reinforcement learning scenarios with both proprioceptive and visual observations.

2781. Scaling FP8 training to trillion-token LLMs

链接：https://iclr.cc/virtual/2025/poster/30430 abstract： We train, for the first time, large language models using FP8 precision on datasets up to 2 trillion tokens --- a 20-fold increase over previous limits. Through these extended training runs, we uncover critical instabilities in FP8 training that were not observable in earlier works with shorter durations. We trace these instabilities to outlier amplification by the SwiGLU activation function. Interestingly, we show, both analytically and empirically, that this

amplification happens only over prolonged training periods, and link it to a SwiGLU weight alignment process. To address this newly identified issue, we introduce Smooth-SwiGLU, a novel modification that ensures stable FP8 training without altering function behavior. We also demonstrate, for the first time, FP8 quantization of both Adam optimizer moments. Combining these innovations, we successfully train a 7B parameter model using FP8 precision on 256 Intel Gaudi2 accelerators, achieving on-par results with the BF16 baseline while delivering up to a $\sim$ 34 % throughput improvement. A reference implementation is supplied in https://github.com/Anonymous1252022/Megatron-DeepSpeed

2782. Enhancing Pre-trained Representation Classifiability can Boost its Interpretability

链接：https://iclr.cc/virtual/2025/poster/30262 abstract： The visual representation of a pre-trained model prioritizes the classifiability on downstream tasks, while the widespread applications for pre-trained visual models have posed new requirements for representation interpretability. However, it remains unclear whether the pre-trained representations can achieve high interpretability and classifiability simultaneously. To answer this question, we quantify the representation interpretability by leveraging its correlation with the ratio of interpretable semantics within the representations. Given the pre-trained representations, only the interpretable semantics can be captured by interpretations, whereas the uninterpretable part leads to information loss. Based on this fact, we propose the Inherent Interpretability Score (IIS) that evaluates the information loss, measures the ratio of interpretable semantics, and quantifies the representation interpretability. In the evaluation of the representation interpretability with different classifiability, we surprisingly discover that the interpretability and classifiability are positively correlated, i.e., representations with higher classifiability provide more interpretable semantics that can be captured in the interpretations. This observation further supports two benefits to the pre-trained representations. First, the classifiability of representations can be further improved by fine-tuning with interpretability maximization. Second, with the classifiability improvement for the representations, we obtain predictions based on their interpretations with less accuracy degradation. The discovered positive correlation and corresponding applications show that practitioners can unify the improvements in interpretability and classifiability for pre-trained vision models. Codes are available at https://github.com/ssfgunner/IIS.

2783. Efficient Multi-agent Offline Coordination via Diffusion-based Trajectory Stitching

链接：https://iclr.cc/virtual/2025/poster/30387 abstract： Learning from offline data without interacting with the environment is a promising way to fully leverage the intelligent decision-making capabilities of multi-agent reinforcement learning (MARL). Previous approaches have primarily focused on developing learning techniques, such as conservative methods tailored to MARL using limited offline data. However, these methods often overlook the temporal relationships across different timesteps and spatial relationships between teammates, resulting in low learning efficiency in imbalanced data scenarios. To comprehensively explore the data structure of MARL and enhance learning efficiency, we propose Multi-Agent offline coordination via Diffusion-based Trajectory Stitching (MADiTS), a novel diffusion-based data augmentation pipeline that systematically generates trajectories by stitching high-quality coordination segments together. MADiTS first generates trajectory segments using a trained diffusion model, followed by applying a bidirectional dynamics constraint to ensure that the trajectories align with environmental dynamics. Additionally, we develop an offline credit assignment technique to identify and optimize the behavior of underperforming agents in the generated segments. This iterative procedure continues until a satisfactory augmented episode trajectory is generated within the predefined limit or is discarded otherwise. Empirical results on imbalanced datasets of multiple benchmarks demonstrate that MADiTS significantly improves MARL performance.

2784. Supervised and Semi-Supervised Diffusion Maps with Label-Driven Diffusion

链接：https://iclr.cc/virtual/2025/poster/30307 abstract： In this paper, we introduce Supervised Diffusion Maps (SDM) and Semi-Supervised Diffusion Maps (SSDM), which transform the well-known unsupervised dimensionality reduction algorithm, Diffusion Maps, into supervised and semi-supervised learning tools. The proposed methods, SDM and SSDM, are based on our new approach that treats the labels as a second view of the data. This unique framework allows us to incorporate ideas from multi-view learning. Specifically, we propose constructing two affinity kernels corresponding to the data and the labels. We then propose a multiplicative interpolation scheme of the two kernels, whose purpose is twofold. First, our scheme extracts the common structure underlying the data and the labels by defining a diffusion process driven by the data and the labels. This label-driven diffusion produces an embedding that emphasizes the properties relevant to the label-related task. Second, the proposed interpolation scheme balances the influence of the two kernels. We show on multiple benchmark datasets that the embedding learned by SDM and SSDM is more effective in downstream regression and classification tasks than existing unsupervised, supervised, and semi-supervised nonlinear dimension reduction methods.

2785. Tree-Wasserstein Distance for High Dimensional Data with a Latent Feature Hierarchy

链接：https://iclr.cc/virtual/2025/poster/28404 abstract： Finding meaningful distances between high-dimensional data samples is an important scientific task. To this end, we propose a new tree-Wasserstein distance (TWD) for high-dimensional data with two key aspects. First, our TWD is specifically designed for data with a latent feature hierarchy, i.e., the features lie in a

hierarchical space, in contrast to the usual focus on embedding samples in hyperbolic space. Second, while the conventional use of TWD is to speed up the computation of the Wasserstein distance, we use its inherent tree as a means to learn the latent feature hierarchy. The key idea of our method is to embed the features into a multi-scale hyperbolic space using diffusion geometry and then present a new tree decoding method by establishing analogies between the hyperbolic embedding and trees. We show that our TWD computed based on data observations provably recovers the TWD defined with the latent feature hierarchy and that its computation is efficient and scalable. We showcase the usefulness of the proposed TWD in applications to word-document and single-cell RNA-sequencing datasets, demonstrating its advantages over existing TWDs and methods based on pre-trained models.

2786. Reassessing How to Compare and Improve the Calibration of Machine Learning Models

链接：https://iclr.cc/virtual/2025/poster/29330 abstract： A machine learning model is calibrated if its predicted probability for an outcome matches the observed frequency for that outcome conditional on the model prediction. This property has become increasingly important as the impact of machine learning models has continued to spread to various domains. As a result, there are now a dizzying number of recent papers on measuring and improving the calibration of (specifically deep learning) models. In this work, we reassess the reporting of calibration metrics in the recent literature. We show that there exist trivial recalibration approaches that can appear seemingly state-of-the-art unless calibration and prediction metrics (i.e. test accuracy) are accompanied by additional generalization metrics such as negative log-likelihood. We then use a calibration-based decomposition of Bregman divergences to develop a new extension to reliability diagrams that jointly visualizes calibration and generalization error, and show how our visualization can be used to detect trade-offs between calibration and generalization. Along the way, we prove novel results regarding the relationship between full calibration error and confidence calibration error for Bregman divergences. We also establish the consistency of the kernel regression estimator for calibration error used in our visualization approach, which generalizes existing consistency results in the literature.

2787. On Stochastic Contextual Bandits with Knapsacks in Small Budget Regime

链接：https://iclr.cc/virtual/2025/poster/30356 abstract： This paper studies stochastic contextual bandits with knapsack constraints (CBwK), where a learner observes a context, takes an action, receives a reward, and incurs a vector of costs at every round. The learner aims to maximize the cumulative rewards across T rounds under the knapsack constraints with an initial budget of B. We study CBwK in the small budget regime where the budget $B = \Omega(\sqrt{T})$and propose an Adaptive and Universal Primal--Dual algorithm (AUPD) that achieves strong regret performance: i) AUPD achieves $\tilde{O}((1 + \frac{\nu^}{\delta b})\sqrt{T})$ regret under the strict feasibility assumption without any prior information, matching the best-known bounds;ii) AUPD achieves $\tilde{O}(\sqrt{T}+ \frac{\nu^}{\sqrt{b}}T^{\frac{3}{4}})$ regret without strict feasibility assumption, which, to the best of our knowledge, is the first result in the literature. Here, the parameter ν^* represents the optimal average reward; $b=B/T$ is the average budget and δb is the feasibility/safety margin.We establish these strong results through the adaptive budget-aware design, which effectively balances reward maximization and budget consumption. We provide a new perspective on analyzing budget consumption using the Lyapunov drift method, along with a refined analysis of its cumulative variance. Our theory is further supported by experiments conducted on a large-scale dataset.

2788. ChemAgent: Self-updating Memories in Large Language Models Improves Chemical Reasoning

链接：https://iclr.cc/virtual/2025/poster/28557 abstract： Chemical reasoning usually involves complex, multi-step processes that demand precise calculations, where even minor errors can lead to cascading failures. Furthermore, large language models (LLMs) encounter difficulties handling domain-specific formulas, executing reasoning steps accurately, and integrating code ef-effectively when tackling chemical reasoning tasks. To address these challenges, we present ChemAgent, a novel framework designed to improve the performance of LLMs through a dynamic, self-updating library. This library is developed by decomposing chemical tasks into sub-tasks and compiling these sub-tasks into a structured collection that can be referenced for future queries. Then, when presented with a new problem, ChemAgent retrieves and refines pertinent information from the library, which we call memory, facilitating effective task decomposition and the generation of solutions. Our method designs three types of memory and a library-enhanced reasoning component, enabling LLMs to improve over time through experience. Experimental results on four chemical reasoning datasets from SciBench demonstrate that ChemAgent achieves performance gains of up to 46% (GPT-4), significantly outperforming existing methods. Our findings suggest substantial potential for future applications, including tasks such as drug discovery and materials science. Our code can be found at https://github.com/gersteinlab/ChemAgent.

2789. For Better or For Worse? Learning Minimum Variance Features With Label Augmentation

链接：https://iclr.cc/virtual/2025/poster/30011 abstract： Data augmentation has been pivotal in successfully training deep learning models on classification tasks over the past decade. An important subclass of data augmentation techniques - which includes both label smoothing and Mixup - involves modifying not only the input data but also the input label during model

training. In this work, we analyze the role played by the label augmentation aspect of such methods. We first prove that linear models on binary classification data trained with label augmentation learn only the minimum variance features in the data, while standard training (which includes weight decay) can learn higher variance features. We then use our techniques to show that even for nonlinear models and general data distributions, the label smoothing and Mixup losses are lower bounded by a function of the model output variance. Lastly, we demonstrate empirically that this aspect of label smoothing and Mixup can be a positive and a negative. On the one hand, we show that the strong performance of label smoothing and Mixup on image classification benchmarks is correlated with learning low variance hidden representations. On the other hand, we show that Mixup and label smoothing can be more susceptible to low variance spurious correlations in the training data.

2790. Audio Large Language Models Can Be Descriptive Speech Quality Evaluators

链接：https://iclr.cc/virtual/2025/poster/29492 abstract： An ideal multimodal agent should be aware of the quality of its input modalities. Recent advances have enabled large language models (LLMs) to incorporate auditory systems for handling various speech-related tasks. However, most audio LLMs remain unaware of the quality of the speech they process. This limitation arises because speech quality evaluation is typically excluded from multi-task training due to the lack of suitable datasets. To address this, we introduce the first natural language-based speech evaluation corpus, generated from authentic human ratings. In addition to the overall Mean Opinion Score (MOS), this corpus offers detailed analysis across multiple dimensions and identifies causes of quality degradation. It also enables descriptive comparisons between two speech samples (A/B tests) with human-like judgment. Leveraging this corpus, we propose an alignment approach with LLM distillation (ALLD) to guide the audio LLM in extracting relevant information from raw speech and generating meaningful responses. Experimental results demonstrate that ALLD outperforms the previous state-of-the-art regression model in MOS prediction, with a mean square error of 0.17 and an A/B test accuracy of 98.6%. Additionally, the generated responses achieve BLEU scores of 25.8 and 30.2 on two tasks, surpassing the capabilities of task-specific models. This work advances the comprehensive perception of speech signals by audio LLMs, contributing to the development of real-world auditory and sensory intelligent agents.

2791. Residual-MPPI: Online Policy Customization for Continuous Control

链接：https://iclr.cc/virtual/2025/poster/28816 abstract： Policies developed through Reinforcement Learning (RL) and Imitation Learning (IL) have shown great potential in continuous control tasks, but real-world applications often require adapting trained policies to unforeseen requirements. While fine-tuning can address such needs, it typically requires additional data and access to the original training metrics and parameters.In contrast, an online planning algorithm, if capable of meeting the additional requirements, can eliminate the necessity for extensive training phases and customize the policy without knowledge of the original training scheme or task. In this work, we propose a generic online planning algorithm for customizing continuous-control policies at the execution time, which we call Residual-MPPI. It can customize a given prior policy on new performance metrics in few-shot and even zero-shot online settings, given access to the prior action distribution alone. Through our experiments, we demonstrate that the proposed Residual-MPPI algorithm can accomplish the few-shot/zero-shot online policy customization task effectively, including customizing the champion-level racing agent, Gran Turismo Sophy (GT Sophy) 1.0, in the challenging car racing scenario, Gran Turismo Sport (GTS) environment. Code for MuJoCo experiments is included in the supplementary and will be open-sourced upon acceptance. Demo videos are available on our website: https://sites.google.com/view/residual-mppi.

2792. ImagineNav: Prompting Vision-Language Models as Embodied Navigator through Scene Imagination

链接：https://iclr.cc/virtual/2025/poster/27914 abstract： Visual navigation is an essential skill for home-assistance robots, providing the object-searching ability to accomplish long-horizon daily tasks. Many recent approaches use Large Language Models (LLMs) for commonsense inference to improve exploration efficiency. However, the planning process of LLMs is limited within texts and it is difficult to represent the spatial occupancy and geometry layout only by texts. Both are important for making rational navigation decisions. In this work, we seek to unleash the spatial perception and planning ability of Vision-Language Models (VLMs), and explore whether the VLM, with only on-board camera captured RGB/RGB-D stream inputs, can efficiently finish the visual navigation tasks in a mapless manner. We achieve this by developing the imagination-powered navigation framework ImagineNav, which imagines the future observation images at valuable robot views and translates the complex navigation planning process into a rather simple best-view image selection problem for VLM. To generate appropriate candidate robot views for imagination, we introduce the Where2Imagine module, which is distilled to align with human navigation habits. Finally, to reach the VLM preferred views, an off-the-shelf point-goal navigation policy is utilized. Empirical experiments on the challenging open-vocabulary object navigation benchmarks demonstrates the superiority of our proposed system.

2793. X-Drive: Cross-modality Consistent Multi-Sensor Data Synthesis for Driving Scenarios

链接：https://iclr.cc/virtual/2025/poster/30182 abstract： Recent advancements have exploited diffusion models for the synthesis of either LiDAR point clouds or camera image data in driving scenarios. Despite their success in modeling single-modality data marginal distribution, there is an under- exploration in the mutual reliance between different modalities to describe com- plex driving scenes. To fill in this gap, we propose a novel framework, X-DRIVE, to model the joint distribution of point

clouds and multi-view images via a dual- branch latent diffusion model architecture. Considering the distinct geometrical spaces of the two modalities, X-DRIVE conditions the synthesis of each modality on the corresponding local regions from the other modality, ensuring better alignment and realism. To further handle the spatial ambiguity during denoising, we design the cross-modality condition module based on epipolar lines to adaptively learn the cross-modality local correspondence. Besides, X-DRIVE allows for controllable generation through multi-level input conditions, including text, bounding box, image, and point clouds. Extensive results demonstrate the high-fidelity synthetic results of X-DRIVE for both point clouds and multi-view images, adhering to input conditions while ensuring reliable cross-modality consistency. Our code will be made publicly available at https://github.com/yichen928/X-Drive.

2794. Self-Attention-Based Contextual Modulation Improves Neural System Identification

链接：https://iclr.cc/virtual/2025/poster/30094 abstract： Convolutional neural networks (CNNs) have been shown to be state-of-the-art models for visual cortical neurons. Cortical neurons in the primary visual cortex are sensitive to contextual information mediated by extensive horizontal and feedback connections. Standard CNNs integrate global contextual information to model contextual modulation via two mechanisms: successive convolutions and a fully connected readout layer. In this paper, we find that self-attention (SA), an implementation of non-local network mechanisms, can improve neural response predictions over parameter-matched CNNs in two key metrics: tuning curve correlation and peak tuning. We introduce peak tuning as a metric to evaluate a model's ability to capture a neuron's top feature preference. We factorize networks to assess each context mechanism, revealing that information in the local receptive field is most important for modeling overall tuning, but surround information is critically necessary for characterizing the tuning peak. We find that self-attention can replace posterior spatial-integration convolutions when learned incrementally, and is further enhanced in the presence of a fully connected readout layer, suggesting that the two context mechanisms are complementary. Finally, we find that decomposing receptive field learning and contextual modulation learning in an incremental manner may be an effective and robust mechanism for learning surround-center interactions.

2795. Restructuring Vector Quantization with the Rotation Trick

链接：https://iclr.cc/virtual/2025/poster/30284 abstract： Vector Quantized Variational AutoEncoders (VQ-VAEs) are designed to compress a continuous input to a discrete latent space and reconstruct it with minimal distortion. They operate by maintaining a set of vectors---often referred to as the codebook---and quantizing each encoder output to the nearest vector in the codebook. However, as vector quantization is non-differentiable, the gradient to the encoder flows around the vector quantization layer rather than through it in a straight-through approximation.This approximation may be undesirable as all information from the vector quantization operation is lost. In this work, we propose a way to propagate gradients through the vector quantization layer of VQ-VAEs. We smoothly transform each encoder output into its corresponding codebook vector via a rotation and rescaling linear transformation that is treated as a constant during backpropagation. As a result, the relative magnitude and angle between encoder output and codebook vector becomes encoded into the gradient as it propagates through the vector quantization layer and back to the encoder.Across 11 different VQ-VAE training paradigms, we find this restructuring improves reconstruction metrics, codebook utilization, and quantization error.

2796. GoodDrag: Towards Good Practices for Drag Editing with Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29395 abstract： In this paper, we introduce GoodDrag, a novel approach to improve the stability and image quality of drag editing. Unlike existing methods that struggle with accumulated perturbations and often result in distortions, GoodDrag introduces an AIDD framework that alternates between drag and denoising operations within the diffusion process, effectively improving the fidelity of the result. We also propose an information-preserving motion supervision operation that maintains the original features of the starting point for precise manipulation and artifact reduction. In addition, we contribute to the benchmarking of drag editing by introducing a new dataset, Drag100, and developing dedicated quality assessment metrics, Dragging Accuracy Index and Gemini Score, utilizing Large Multimodal Models. Extensive experiments demonstrate that the proposed GoodDrag compares favorably against the state-of-the-art approaches both qualitatively and quantitatively. The source code and data are available at https://gooddrag.github.io.

2797. RM-Bench: Benchmarking Reward Models of Language Models with Subtlety and Style

链接：https://iclr.cc/virtual/2025/poster/29702 abstract： Reward models are critical in techniques like Reinforcement Learning from Human Feedback (RLHF) and Inference Scaling Laws, where they guide language model alignment and select optimal responses. Despite their importance, existing reward model benchmarks often evaluate models by asking them to distinguish between responses generated by models of varying power. However, this approach fails to assess reward models on subtle but critical content changes and variations in style, resulting in a low correlation with policy model performance.To this end, we introduce RM-Bench, a novel benchmark designed to evaluate reward models based on their sensitivity to subtle content differences and resistance to style biases. Extensive experiments demonstrate that RM-Bench strongly correlates with policy model performance, making it a reliable reference for selecting reward models to align language models effectively.We evaluate

nearly 40 reward models on RM-Bench. Our results reveal that even state-of-the-art models achieve an average performance of only 46.6%, which falls short of random-level accuracy (50%) when faced with style bias interference.These findings highlight the significant room for improvement in current reward models.

2798. Analyzing Neural Scaling Laws in Two-Layer Networks with Power-Law Data Spectra

链接：https://iclr.cc/virtual/2025/poster/27863 abstract： Neural scaling laws describe how the performance of deep neural networks scales with key factors such as training data size, model complexity, and training time, often following power-law behaviors over multiple orders of magnitude. Despite their empirical observation, the theoretical understanding of these scaling laws remains limited. In this work, we employ techniques from statistical mechanics to analyze one-pass stochastic gradient descent within a student-teacher framework, where both the student and teacher are two-layer neural networks. Our study primarily focuses on the generalization error and its behavior in response to data covariance matrices that exhibit power-law spectra.For linear activation functions, we derive analytical expressions for the generalization error, exploring different learning regimes and identifying conditions under which power-law scaling emerges. Additionally, we extend our analysis to non-linear activation functions in the feature learning regime, investigating how power-law spectra in the data covariance matrix impact learning dynamics. Importantly, we find that the length of the symmetric plateau depends on the number of distinct eigenvalues of the data covariance matrix and the number of hidden units, demonstrating how these plateaus behave under various configurations. In addition, our results reveal a transition from exponential to power-law convergence in the specialized phase when the data covariance matrix possesses a power-law spectrum. This work contributes to the theoretical understanding of neural scaling laws and provides insights into optimizing learning performance in practical scenarios involving complex data structures.

2799. Distribution Backtracking Builds A Faster Convergence Trajectory for Diffusion Distillation

链接：https://iclr.cc/virtual/2025/poster/31101 abstract：

2800. FaceShot: Bring Any Character into Life

链接：https://iclr.cc/virtual/2025/poster/28365 abstract： In this paper, we present FaceShot, a novel training-free portrait animation framework designed to bring any character into life from any driven video without fine-tuning or retraining.We achieve this by offering precise and robust reposed landmark sequences from an appearance-guided landmark matching module and a coordinate-based landmark retargeting module.Together, these components harness the robust semantic correspondences of latent diffusion models to produce facial motion sequence across a wide range of character types.After that, we input the landmark sequences into a pre-trained landmark-driven animation model to generate animated video.With this powerful generalization capability, FaceShot can significantly extend the application of portrait animation by breaking the limitation of realistic portrait landmark detection for any stylized character and driven video.Also, FaceShot is compatible with any landmark-driven animation model, significantly improving overall performance.Extensive experiments on our newly constructed character benchmark CharacBench confirm that FaceShot consistently surpasses state-of-the-art (SOTA) approaches across any character domain.More results are available at our project website https://faceshot2024.github.io/faceshot/.