1. Universal Image Restoration Pre-training via Degradation Classification

链接：https://iclr.cc/virtual/2025/poster/29746 abstract： This paper proposes the Degradation Classification Pre-Training (DCPT), which enables models to learn how to classify the degradation type of input images for universal image restoration pre-training. Unlike the existing self-supervised pre-training methods, DCPT utilizes the degradation type of the input image as an extremely weak supervision, which can be effortlessly obtained, even intrinsic in all image restoration datasets. DCPT comprises two primary stages. Initially, image features are extracted from the encoder. Subsequently, a lightweight decoder, such as ResNet18, is leveraged to classify the degradation type of the input image solely based on the features extracted in the first stage, without utilizing the input image. The encoder is pre-trained with a straightforward yet potent DCPT, which is used to address universal image restoration and achieve outstanding performance. Following DCPT, both convolutional neural networks (CNNs) and transformers demonstrate performance improvements, with gains of up to 2.55 dB in the 10D all-in-one restoration task and 6.53 dB in the mixed degradation scenarios. Moreover, previous self-supervised pretraining methods, such as masked image modeling, discard the decoder after pre-training, while our DCPT utilizes the pre-trained parameters more effectively. This superiority arises from the degradation classifier acquired during DCPT, which facilitates transfer learning between models of identical architecture trained on diverse degradation types. Source code and models are available at \url{https://github.com/MILab-PKU/dcpt}.

2. On the Benefits of Attribute-Driven Graph Domain Adaptation

链接：https://iclr.cc/virtual/2025/poster/28080 abstract： Graph Domain Adaptation (GDA) addresses a pressing challenge in cross-network learning, particularly pertinent due to the absence of labeled data in real-world graph datasets. Recent studies attempted to learn domain invariant representations by eliminating structural shifts between graphs. In this work, we show that existing methodologies have overlooked the significance of the graph node attribute, a pivotal factor for graph domain alignment. Specifically, we first reveal the impact of node attributes for GDA by theoretically proving that in addition to the graph structural divergence between the domains, the node attribute discrepancy also plays a critical role in GDA. Moreover, we also empirically show that the attribute shift is more substantial than the topology shift, which further underscore the importance of node attribute alignment in GDA. Inspired by this finding, a novel cross-channel module is developed to fuse and align both views between the source and target graphs for GDA. Experimental results on a variety of benchmark verify the effectiveness of our method.

3. Port-Hamiltonian Architectural Bias for Long-Range Propagation in Deep Graph Networks

链接：https://iclr.cc/virtual/2025/poster/31273 abstract： The dynamics of information diffusion within graphs is a critical open issue that heavily influences graph representation learning, especially when considering long-range propagation. This calls for principled approaches that control and regulate the degree of propagation and dissipation of information throughout the neural flow. Motivated by this, we introduce port-Hamiltonian Deep Graph Networks, a novel framework that models neural information flow in graphs by building on the laws of conservation of Hamiltonian dynamical systems. We reconcile under a single theoretical and practical framework both non-dissipative long-range propagation and non-conservative behaviors, introducing tools from mechanical systems to gauge the equilibrium between the two components. Our approach can be applied to general message-passing architectures, and it provides theoretical guarantees on information conservation in time. Empirical results prove the effectiveness of our port-Hamiltonian scheme in pushing simple graph convolutional architectures to state-of-the-art performance in long-range benchmarks.

4. SORRY-Bench: Systematically Evaluating Large Language Model Safety Refusal

链接：https://iclr.cc/virtual/2025/poster/29252 abstract： Evaluating aligned large language models' (LLMs) ability to recognize and reject unsafe user requests is crucial for safe, policy-compliant deployments. Existing evaluation efforts, however, face three limitations that we address with SORRY-Bench, our proposed benchmark. First, existing methods often use coarse-grained taxonomies of unsafe topics, and are over-representing some fine-grained topics. For example, among the ten existing datasets that we evaluated, tests for refusals of self-harm instructions are over 3x less represented than tests for fraudulent activities. SORRY-Bench improves on this by using a fine-grained taxonomy of 44 potentially unsafe topics, and 440 class-balanced unsafe instructions, compiled through human-in-the-loop methods. Second, evaluations often overlook the linguistic formatting of prompts, like different languages, dialects, and more --- which are only implicitly considered in many evaluations. We supplement SORRY-bench with 20 diverse linguistic augmentations to systematically examine these effects. Third, existing evaluations rely on large LLMs (e.g., GPT-4) for evaluation, which can be computationally expensive. We investigate design choices for creating a fast, accurate automated safety evaluator. By collecting 7K+ human annotations and conducting a meta-evaluation of diverse LLM-as-a-judge designs, we show that fine-tuned 7B LLMs can achieve accuracy comparable to GPT-4 scale LLMs, with lower computational cost. Putting these together, we evaluate over 50 proprietary and open-weight LLMs on SORRY-Bench, analyzing their distinctive safety refusal behaviors. We hope our effort provides a building block for systematic evaluations of LLMs' safety refusal capabilities, in a balanced, granular, and efficient manner. Benchmark demo, data, code, and models are available through https://sorry-bench.github.io.

5. Framer: Interactive Frame Interpolation

链接：https://iclr.cc/virtual/2025/poster/29976 abstract： We propose Framer for interactive frame interpolation, which targets producing smoothly transitioning frames between two images as per user creativity. Concretely, besides taking the start and end frames as inputs, our approach supports customizing the transition process by tailoring the trajectory of some selected keypoints. Such a design enjoys two clear benefits. First, incorporating human interaction mitigates the issue arising from numerous possibilities of transforming one image to another, and in turn enables finer control of local motions. Second, as the most basic form of interaction, keypoints help establish the correspondence across frames, enhancing the model to handle challenging cases (e.g., objects on the start and end frames are of different shapes and styles). It is noteworthy that our system also offers an "autopilot" mode, where we introduce a module to estimate the keypoints and refine the trajectory automatically, to simplify the usage in practice. Extensive experimental results demonstrate the appealing performance of Framer on various applications, such as image morphing, time-lapse video generation, cartoon interpolation, etc. The code, model, and interface are publicly accessible at https://github.com/aim-uofa/Framer.

6. Beyond Next Token Prediction: Patch-Level Training for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28998 abstract： The prohibitive training costs of Large Language Models (LLMs) have emerged as a significant bottleneck in the development of next-generation LLMs. In this paper, we show that it is possible to significantly reduce the training costs of LLMs without sacrificing their performance. Specifically, we introduce patch-level training for LLMs, in which multiple tokens are aggregated into a unit of higher information density, referred to as a `patch', to serve as the fundamental text unit for training LLMs. During patch-level training, we feed the language model shorter sequences of patches and train it to predict the next patch, thereby processing the majority of the training data at a significantly reduced cost. Following this, the model continues token-level training on the remaining training data to align with the inference mode. Experiments on a diverse range of models (370M-2.7B parameters) demonstrate that patch-level training can reduce the overall training costs to 0.5$\times$, without compromising the model performance compared to token-level training. Source code: \url{https://github.com/shaochenze/PatchTrain}.

7. CTSyn: A Foundation Model for Cross Tabular Data Generation

链接：https://iclr.cc/virtual/2025/poster/29585 abstract： Generative Foundation Models (GFMs) have achieved remarkable success in producing high-quality synthetic data for images and text. However, their application to tabular data presents significant challenges due to the heterogeneous nature of table features. Current cross-table learning frameworks struggle because they lack a generative model backbone and an effective mechanism to decode heterogeneous feature values. To address these challenges, we propose the Cross-Table Synthesizer (CTSyn), a diffusion-based generative foundation model for tabular data generation. CTSyn comprises two key components. The first is an autoencoder network that consolidates diverse tables into a unified latent space. It dynamically reconstructs table values using a table schema embedding, allowing adaptation to heterogeneous datasets. The second is a conditional latent diffusion model that generates samples from the learned latent space, conditioned on the table schema. Through large-scale pre-training, CTSyn outperforms existing table synthesizers on standard benchmarks in both utility and diversity. These results position CTSyn as a promising framework for synthetic table generation and lay the groundwork for developing large-scale tabular foundation models.

8. Aligned Better, Listen Better for Audio-Visual Large Language Models

链接：https://iclr.cc/virtual/2025/poster/31200 abstract： Audio is essential for multimodal video understanding. On the one hand, video inherently contains audio, which supplies complementary information to vision. Besides, video large language models (Video-LLMs) can encounter many audio-centric settings. However, existing Video-LLMs and Audio-Visual Large Language Models (AV-LLMs) exhibit deficiencies in exploiting audio information, leading to weak understanding and hallucinations. To solve the issues, we delve into the model architecture and dataset. (1) From the architectural perspective, we propose a fine-grained AV-LLM, namely Dolphin. The concurrent alignment of audio and visual modalities in both temporal and spatial dimensions ensures a comprehensive and accurate understanding of videos. Specifically, we devise an audio-visual multi-scale adapter for multi-scale information aggregation, which achieves spatial alignment. For temporal alignment, we propose audio-visual interleaved merging. (2) From the dataset perspective, we curate an audio-visual caption \& instruction-tuning dataset, called AVU. It comprises 5.2 million diverse, open-ended data tuples (video, audio, question, answer) and introduces a novel data partitioning strategy. Extensive experiments show our model not only achieves remarkable performance in audio-visual understanding, but also mitigates potential hallucinations.

9. MiniPLM: Knowledge Distillation for Pre-training Language Models

链接：https://iclr.cc/virtual/2025/poster/28062 abstract： Knowledge distillation (KD) is widely used to train small, high-performing student language models (LMs) using large teacher LMs. While effective in fine-tuning, KD during pre-training faces efficiency, flexibility, and effectiveness issues. Existing methods either incur high computational costs due to online teacher inference, require tokenization matching between teacher and student LMs, or risk losing the difficulty and diversity of the teacher-generated training data.In this work, we propose MiniPLM, a KD framework for pre-training LMs by refining the training data distribution with the teacher LM's knowledge.For efficiency, MiniPLM performs offline teacher inference, allowing KD for

multiple student LMs without adding training costs.For flexibility, MiniPLM operates solely on the training corpus, enabling KD across model families.For effectiveness, MiniPLM leverages the differences between large and small LMs to enhance the training data difficulty and diversity, helping student LMs acquire versatile and sophisticated knowledge.Extensive experiments demonstrate that MiniPLM boosts the student LMs' performance on 9 common downstream tasks, improves language modeling capabilities, and reduces pre-training computation. The benefit of MiniPLM extends to larger training scales, evidenced by the scaling curve extrapolation.Further analysis reveals that MiniPLM supports KD across model families and enhances the pre-training data utilization. Our code, data, and models can be found at https://github.com/thu-coai/MiniPLM.

10. From an LLM Swarm to a PDDL-empowered Hive: Planning Self-executed Instructions in a Multi-modal Jungle

链接：https://iclr.cc/virtual/2025/poster/29706 abstract： In response to the call for agent-based solutions that leverage the ever-increasing capabilities of the deep models' ecosystem, we introduce a comprehensive solution for selecting appropriate models and subsequently planning a set of atomic actions to satisfy the end-users' instructions.Our system, Hive, operates over sets of models and, upon receiving natural language instructions, schedules and executes, explainable plans of atomic actions. These actions can involve one or more of the available models to achieve the overall task, while respecting end-users specific constraints. Hive is able to plan complex chains of actions while guaranteeing explainability, using an LLM-based formal logic backbone empowered by PDDL operations. We introduce the MuSE benchmark in order to offer a comprehensive evaluation of the multi-modal capabilities of agent systems. Our findings show that our framework redefines the state-of-the-art for task selection, outperforming other competing systems that plan operations across multiple models while offering transparency guarantees while fully adhering to user constraints.

11. Recite, Reconstruct, Recollect: Memorization in LMs as a Multifaceted Phenomenon

链接：https://iclr.cc/virtual/2025/poster/31090 abstract： Memorization in language models is typically treated as a homogenous phenomenon, neglecting the specifics of the memorized data. We instead model memorization as the effect of a set of complex factors that describe each sample and relate it to the model and corpus. To build intuition around these factors, we break memorization down into a taxonomy: recitation of highly duplicated sequences, reconstruction of inherently predictable sequences, and recollection of sequences that are neither. We demonstrate the usefulness of our taxonomy by using it to construct a predictive model for memorization. By analyzing dependencies and inspecting the weights of the predictive model, we find that different factors have different influences on the likelihood of memorization depending on the taxonomic category.

12. DelTA: An Online Document-Level Translation Agent Based on Multi-Level Memory

链接：https://iclr.cc/virtual/2025/poster/28742 abstract： Large language models (LLMs) have achieved reasonable quality improvements in machine translation (MT).However, most current research on MT-LLMs still faces significant challenges in maintaining translation consistency and accuracy when processing entire documents.In this paper, we introduce DelTA, a Document-levEL Translation Agent designed to overcome these limitations.DelTA features a multi-level memory structure that stores information across various granularities and spans, including Proper Noun Records, Bilingual Summary, Long-Term Memory, and Short-Term Memory, which are continuously retrieved and updated by auxiliary LLM-based components.Experimental results indicate that DelTA significantly outperforms strong baselines in terms of translation consistency and quality across four open/closed-source LLMs and two representative document translation datasets, achieving an increase in consistency scores by up to 4.58 percentage points and in COMET scores by up to 3.16 points on average.DelTA employs a sentence-by-sentence translation strategy, ensuring no sentence omissions and offering a memory-efficient solution compared to the mainstream method.Furthermore, DelTA improves pronoun and context-dependent translation accuracy, and the summary component of the agent also shows promise as a tool for query-based summarization tasks.The code and data of our approach are released at https://github.com/YutongWang1216/DocMTAgent.

13. Discrete Codebook World Models for Continuous Control

链接：https://iclr.cc/virtual/2025/poster/28506 abstract： In reinforcement learning (RL), world models serve as internal simulators, enabling agents to predict environment dynamics and future outcomes in order to make informed decisions. While previous approaches leveraging discrete latent spaces, such as DreamerV3, have demonstrated strong performance in discrete action settings and visual control tasks, their comparative performance in state-based continuous control remains underexplored. In contrast, methods with continuous latent spaces, such as TD-MPC2, have shown notable success in state-based continuous control benchmarks. In this paper, we demonstrate that modeling discrete latent states has benefits over continuous latent states and that discrete codebook encodings are more effective representations for continuous control, compared to alternative encodings, such as one-hot and label-based encodings. Based on these insights, we introduce DCWM: Discrete Codebook World Model, a self-supervised world model with a discrete and stochastic latent space, where latent states are codes from a codebook. We combine DCWM with decision-time planning to get our model-based RL algorithm, named DC-MPC: Discrete Codebook Model Predictive Control, which performs competitively against recent state-of-the-art algorithms, including TD-MPC2 and DreamerV3, on continuous control benchmarks.

14. The 3D-PC: a benchmark for visual perspective taking in humans and machines

链接：https://iclr.cc/virtual/2025/poster/29481 abstract： Visual perspective taking (VPT) is the ability to perceive and reason about the perspectives of others. It is an essential feature of human intelligence, which develops over the first decade of life and requires an ability to process the 3D structure of visual scenes. A growing number of reports have indicated that deep neural networks (DNNs) become capable of analyzing 3D scenes after training on large image datasets. We investigated if this emergent ability for 3D analysis in DNNs is sufficient for VPT with the 3D perception challenge (3D-PC): a novel benchmark for 3D perception in humans and DNNs. The 3D-PC is comprised of three 3D-analysis tasks posed within natural scene images: (i.) a simple test of object depth order, (ii.) a basic VPT task (VPT-basic), and (iii.) a more challenging version of VPT (VPT-perturb) designed to limit the effectiveness of "shortcut" visual strategies. We tested human participants (N=33) and linearly probed or text-prompted over 300 DNNs on the challenge and found that nearly all of the DNNs approached or exceeded human accuracy in analyzing object depth order. Surprisingly, DNN accuracy on this task correlated with their object recognition performance. In contrast, there was an extraordinary gap between DNNs and humans on VPT-basic. Humans were nearly perfect, whereas most DNNs were near chance. Fine-tuning DNNs on VPT-basic brought them close to human performance, but they, unlike humans, dropped back to chance when tested on VPT-perturb. Our challenge demonstrates that the training routines and architectures of today's DNNs are well-suited for learning basic 3D properties of scenes and objects but are ill-suited for reasoning about these properties like humans do. We release our 3D-PC datasets and code to help bridge this gap in 3D perception between humans and machines.

15. Contextual Self-paced Learning for Weakly Supervised Spatio-Temporal Video Grounding

链接：https://iclr.cc/virtual/2025/poster/27737 abstract： In this work, we focus on Weakly Supervised Spatio-Temporal Video Grounding (WSTVG). It is a multimodal task aimed at localizing specific subjects spatio-temporally based on textual queries without bounding box supervision. Motivated by recent advancements in multi-modal foundation models for grounding tasks, we first explore the potential of state-of-the-art object detection models for WSTVG. Despite their robust zero-shot capabilities, our adaptation reveals significant limitations, including inconsistent temporal predictions, inadequate understanding of complex queries, and challenges in adapting to difficult scenarios. We propose CoSPaL (Contextual Self-Paced Learning), a novel approach which is designed to overcome these limitations. CoSPaL integrates three core components: (1) Tubelet Phrase Grounding (TPG), which introduces spatio-temporal prediction by linking textual queries to tubelets; (2) Contextual Referral Grounding (CRG), which improves comprehension of complex queries by extracting contextual information to refine object identification over time; and (3) Self-Paced Scene Understanding (SPS), a training paradigm that progressively increases task difficulty, enabling the model to adapt to complex scenarios by transitioning from coarse to fine-grained understanding.

16. Faster Cascades via Speculative Decoding

链接：https://iclr.cc/virtual/2025/poster/27888 abstract： Cascades and speculative decoding are two common approaches to improving language models' inference efficiency. Both approaches interleave two models, but via fundamentally distinct mechanisms: deferral rule that invokes the larger model only for “hard” inputs, while speculative decoding uses speculative execution to primarily invoke the larger model in parallel scoring mode. These mechanisms offer different benefits: empirically, cascades offer compelling cost-quality trade-offs, often even outperforming the large model; speculative cascades offer impressive speed-ups, while guaranteeing quality-neutrality. In this paper, we leverage the best of both these approaches by designing new speculative cascading techniques that implement their deferral rule through speculative execution. We characterize the optimal deferral rule for our speculative cascades, and employ a plug-in approximation to the optimal rule. Experiments with Gemma and T5 models on a range of language benchmarks show that our approach yields better cost quality trade-offs than cascading and speculative decoding baselines.

17. miniCTX: Neural Theorem Proving with (Long-)Contexts

链接：https://iclr.cc/virtual/2025/poster/30064 abstract： Real-world formal theorem proving often depends on a wealth of context, including definitions, lemmas, comments, file structure, and other information. We introduce $\texttt{miniCTX}$, which tests a model's ability to prove formal mathematical theorems that depend on new context that is not seen during training. $\texttt{miniCTX}$ contains theorems sourced from real Lean projects and textbooks, each associated with a context that can span tens of thousands of tokens. Models are tasked with proving a theorem given access to code from the theorem's repository, which contains context that is needed for the proof. As a baseline for $\texttt{miniCTX}$, we tested fine-tuning and prompting methods that condition theorem proving on preceding context. Both approaches substantially outperform traditional methods that rely solely on state information. We found that this ability to use context is not captured by previous benchmarks such as $\texttt{miniF2F}$. Alongside $\texttt{miniCTX}$, we offer $\texttt{ntp-toolkit}$ for automatically extracting and annotating theorem proving data, making it easy to add new projects into $\texttt{miniCTX}$ to ensure that contexts are not seen during training. $\texttt{miniCTX}$ offers a challenging and realistic evaluation of neural theorem provers.

18. Directional Gradient Projection for Robust Fine-Tuning of Foundation Models

链接：https://iclr.cc/virtual/2025/poster/28802 abstract： Robust fine-tuning aims to adapt large foundation models to downstream tasks while preserving their robustness to distribution shifts. Existing methods primarily focus on constraining and projecting current model towards the pre-trained initialization based on the magnitudes between fine-tuned and pre-trained weights, which often require extensive hyper-parameter tuning and can sometimes result in underfitting. In this work, we propose $\textbf{Di}$rectional $\textbf{Gra}$dient $\textbf{P}$rojection (DiGraP), a novel layer-wise trainable method that incorporates directional information from gradients to bridge regularization and multi-objective optimization. Besides demonstrating our method on image classification, as another contribution we generalize this area to the multi-modal evaluation settings for robust fine-tuning. Specifically, we first bridge the uni-modal and multi-modal gap by performing analysis on Image Classification reformulated Visual Question Answering (VQA) benchmarks and further categorize ten out-of-distribution (OOD) VQA datasets by distribution shift types and degree (i.e. near versus far OOD). Experimental results show that DiGraP consistently outperforms existing baselines across Image Classfication and VQA tasks with discriminative and generative backbones, improving both in-distribution (ID) generalization and OOD robustness.

19. Towards a Unified and Verified Understanding of Group-Operation Networks

链接：https://iclr.cc/virtual/2025/poster/30732 abstract： A recent line of work in mechanistic interpretability has focused on reverse-engineering the computation performed by neural networks trained on the binary operation of finite groups. We investigate the internals of one-hidden-layer neural networks trained on this task, revealing previously unidentified structure and producing a more complete description of such models in a step towards unifying the explanations of previous works (Chughtai et al., 2023; Stander et al., 2024). Notably, these models approximate equivariance in each input argument. We verify that our explanation applies to a large fraction of networks trained on this task by translating it into a compact proof of model performance, a quantitative evaluation of the extent to which we faithfully and concisely explain model internals. In the main text, we focus on the symmetric group S5. For models trained on this group, our explanation yields a guarantee of model accuracy that runs 3x faster than brute force and gives a >=95% accuracy bound for 45% of the models we trained. We were unable to obtain nontrivial non-vacuous accuracy bounds using only explanations from previous works.

20. Scaling Diffusion Language Models via Adaptation from Autoregressive Models

链接：https://iclr.cc/virtual/2025/poster/28659 abstract： Diffusion Language Models (DLMs) have emerged as a promising new paradigm for text generative modeling, potentially addressing limitations of autoregressive (AR) models. However, current DLMs have been studied at a smaller scale compared to their AR counterparts and lack fair comparison on language modeling benchmarks. Additionally, training diffusion models from scratch at scale remains challenging. Given the prevalence of open-source AR language models, we propose adapting these models to build text diffusion models. We demonstrate connections between AR and diffusion modeling objectives and introduce a simple continual pre-training approach for training diffusion models. Through systematic evaluation on language modeling, reasoning, and commonsense benchmarks, we show that we can convert AR models ranging from 127M to 7B parameters (GPT2 and LLaMA) into diffusion models DiffuGPT and DiffuLLaMA, using less than 200B tokens for training. Our experimental results reveal that these models outperform earlier DLMs and are competitive with their AR counterparts. We release a suite of DLMs (127M-355M-7B) capable of generating fluent text, performing in-context learning, filling in the middle without prompt re-ordering, and following instructions.

21. Multi-domain Distribution Learning for De Novo Drug Design

链接：https://iclr.cc/virtual/2025/poster/28837 abstract： We introduce DrugFlow, a generative model for structure-based drug design that integrates continuous flow matching with discrete Markov bridges, demonstrating state-of-the-art performance in learning chemical, geometric, and physical aspects of three-dimensional protein-ligand data. We endow DrugFlow with an uncertainty estimate that is able to detect out-of-distribution samples. To further enhance the sampling process towards distribution regions with desirable metric values, we propose a joint preference alignment scheme applicable to both flow matching and Markov bridge frameworks. Furthermore, we extend our model to also explore the conformational landscape of the protein by jointly sampling side chain angles and molecules.

22. NeSyC: A Neuro-symbolic Continual Learner For Complex Embodied Tasks in Open Domains

链接：https://iclr.cc/virtual/2025/poster/29399 abstract： We explore neuro-symbolic approaches to generalize actionable knowledge, enabling embodied agents to tackle complex tasks more effectively in open-domain environments. A key challenge for embodied agents is the generalization of knowledge across diverse environments and situations, as limited experiences often confine them to their prior knowledge. To address this issue, we introduce a novel framework, NeSyC, a neuro-symbolic continual learner that emulates the hypothetico-deductive model by continually formulating and validating knowledge from limited experiences through the combined use of Large Language Models (LLMs) and symbolic tools. Specifically, we devise a contrastive generality improvement scheme within NeSyC, which iteratively generates hypotheses using LLMs and conducts contrastive validation via symbolic tools. This scheme reinforces the justification for admissible actions while minimizing the inference of inadmissible ones. Additionally, we incorporate a memory-based monitoring scheme that efficiently detects action

errors and triggers the knowledge refinement process across domains. Experiments conducted on diverse embodied task benchmarks—including ALFWorld, VirtualHome, Minecraft, RLBench, and a real-world robotic scenario—demonstrate that NeSyC is highly effective in solving complex embodied tasks across a range of open-domain environments.

23. Toward Exploratory Inverse Constraint Inference with Generative Diffusion Verifiers

链接：https://iclr.cc/virtual/2025/poster/31251 abstract： An important prerequisite for safe control is aligning the policy with the underlying constraints in the environment. In many real-world applications, due to the difficulty of manually specifying these constraints, existing works have proposed recovering constraints from expert demonstrations by solving the Inverse Constraint Learning (ICL) problem. However, ICL is inherently ill-posed, as multiple constraints can equivalently explain the experts' preferences, making the optimal solutions not uniquely identifiable. In this work, instead of focusing solely on a single constraint, we propose the novel approach of Exploratory ICL (ExICL). The goal of ExICL is to recover a diverse set of feasible constraints, thereby providing practitioners the flexibility to select the most appropriate constraint based on the practical needs of deployment. To achieve this goal, we design a generative diffusion verifier that guides the trajectory generation process using the probabilistic representation of an optimal constrained policy. By comparing these decisions with those made by expert agents, we can efficiently verify a candidate constraint. Driven by the verification feedback, ExICL implements an exploratory constraint update mechanism that strategically facilitates diversity within the collection of feasible constraints. Our empirical results demonstrate that ExICL can seamlessly and reliably generalize across different tasks and environments. The code is available at https://github.com/ZhaoRunyi/ExICL.

24. Federated Domain Generalization with Data-free On-server Matching Gradient

链接：https://iclr.cc/virtual/2025/poster/30766 abstract： Domain Generalization (DG) aims to learn from multiple known source domains a model that can generalize well to unknown target domains. One of the key approaches in DG is training an encoder which generates domain-invariant representations. However, this approach is not applicable in Federated Domain Generalization (FDG), where data from various domains are distributed across different clients. In this paper, we introduce a novel approach, dubbed Federated Learning via On-server Matching Gradient (FedOMG), which can efficiently leverage domain information from distributed domains. Specifically, we utilize the local gradients as information about the distributed models to find an invariant gradient direction across all domains through gradient inner product maximization. The advantages are two-fold: 1) FedOMG can aggregate the characteristics of distributed models on the centralized server without incurring any additional communication cost, and 2) FedOMG is orthogonal to many existing FL/FDG methods, allowing for additional performance improvements by being seamlessly integrated with them. Extensive experimental evaluations on various settings demonstrate the robustness of FedOMG compared to other FL/FDG baselines. Our method outperforms recent SOTA baselines on four FL benchmark datasets (MNIST, EMNIST, CIFAR-10, and CIFAR-100), and three FDG benchmark datasets (PACS, VLCS, and OfficeHome). The reproducible code is publicly available~\footnote[1]{\url{https://github.com/skydvn/fedomg}}.

25. Context-Parametric Inversion: Why Instruction Finetuning May Not Actually Improve Context Reliance

链接：https://iclr.cc/virtual/2025/poster/29609 abstract： Large Language Model's are instruction-finetuned to enhance their ability to follow user instructions and better comprehend input context. Still, they often struggle to follow the input context, especially when it contradicts model's parametric knowledge. This manifests as various failures, such as hallucinations where a model inserts outdated or unwarranted facts into its response. In this work, we observe an intriguing phenomenon: the context reliance of the model decreases as instruction finetuning progresses, $\textit{despite an initial expected increase}$. We call this phenomenon as the $\textbf{context-parametric inversion}$. This is surprising, as one would expect instruction tuning to improve the model's ability to follow input instructions. We observe this behavior on multiple general purpose instruction tuning datasets such as TULU, Alpaca and Ultrachat, across multiple model families like Llama, Mistral and Pythia. We perform various controlled studies to eliminate some simple hypothesis for this observed behavior and isolate what datapoints cause this counter-intuitive behavior. We then analyze the phenomenon theoretically, to explain why context reliance varies across the trajectory of finetuning. We tie the observed context-parametric inversion to the properties of the finetuning data, which provides us with some potential mitigation strategies that provide limited but insightful gains.

26. Understanding Optimization in Deep Learning with Central Flows

链接：https://iclr.cc/virtual/2025/poster/28135 abstract： Optimization in deep learning remains poorly understood. A key difficulty is that optimizers exhibit complex oscillatory dynamics, referred to as "edge of stability," which cannot be captured by traditional optimization theory. In this paper, we show that the path taken by an oscillatory optimizer can often be captured by a central flow: a differential equation which directly models the time-averaged (i.e. smoothed) optimization trajectory. We empirically show that these central flows can predict long-term optimization trajectories for generic neural networks with a high degree of numerical accuracy. By interpreting these flows, we are able to understand how gradient descent makes progress even as the loss sometimes goes up; how adaptive optimizers ``adapt'' to the local loss landscape; and how adaptive optimizers implicitly seek out regions of weight space where they can take larger steps. These insights (and others) are not apparent from

the optimizers' update rules, but are revealed by the central flows. Therefore, we believe that central flows constitute a promising tool for reasoning about optimization in deep learning.

27. Inference Optimal VLMs Need Fewer Visual Tokens and More Parameters

链接：https://iclr.cc/virtual/2025/poster/30880 abstract： Vision Language Models (VLMs) have demonstrated strong capabilities across various visual understanding and reasoning tasks, driven by incorporating image representations into the token inputs of Large Language Models (LLMs). However, their real-world deployment is often constrained by high latency during inference due to the substantial compute required by the LLM to process the large number of input tokens, predominantly arising from the image. To reduce inference costs, one can either downsize the LLM or reduce the number of input tokens needed to represent the image, the latter of which has been the focus of many recent efforts around token compression. However, it is unclear what the optimal trade-off is given a fixed inference budget. We first characterize this optimal trade-off between the number of visual tokens and LLM parameters by establishing scaling laws that capture variations in performance with these two factors. Our results reveal a surprising trend: for visual reasoning tasks, the inference-optimal behavior in VLMs is achieved by using the largest LLM that fits within the inference budget while minimizing visual token count - often to a single token. While the token reduction literature has mainly focused on maintaining base model performance by modestly reducing the token count (e.g., $5-10\times$), our results indicate that the compute-optimal inference regime requires operating under even higher token compression ratios. Based on these insights, we take the first steps toward designing token compression algorithms tailored for high-compression settings, utilizing prompt-based compression of tokens. Our work underscores the performance and efficiency benefits of operating in low visual token regimes and the importance of developing tailored token reduction algorithms for such conditions.

28. StochSync: Stochastic Diffusion Synchronization for Image Generation in Arbitrary Spaces

链接：https://iclr.cc/virtual/2025/poster/29307 abstract： We propose a zero-shot method for generating images in arbitrary spaces (e.g., a sphere for 360◦ panoramas and a mesh surface for texture) using a pretrained image diffusion model. The zero-shot generation of various visual content using a pretrained image diffusion model has been explored mainly in two directions. First, Diffusion Synchronization–performing reverse diffusion processes jointly across different projected spaces while synchronizing them in the target space–generates high-quality outputs when enough conditioning is provided, but it struggles in its absence. Second, Score Distillation Sampling–gradually updating the target space data through gradient descent–results in better coherence but often lacks detail. In this paper, we reveal for the first time the interconnection between these two methods while highlighting their differences. To this end, we propose StochSync, a novel approach that combines the strengths of both, enabling effective performance with weak conditioning. Our experiments demonstrate that StochSync provides the best performance in 360◦ panorama generation (where image conditioning is not given), outperforming previous finetuning-based methods, and also delivers comparable results in 3D mesh texturing (where depth conditioning is provided) with previous methods.

29. Data Unlearning in Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29573 abstract： Recent work has shown that diffusion models memorize and reproduce training data examples. At the same time, large copyright lawsuits and legislation such as GDPR have highlighted the need for erasing datapoints from diffusion models. However, retraining from scratch is often too expensive. This motivates the setting of data unlearning, i.e., the study of efficient techniques for unlearning specific datapoints from the training set. Existing concept unlearning techniques require an anchor prompt/class/distribution to guide unlearning, which is not available in the data unlearning setting. General-purpose machine unlearning techniques were found to be either unstable or failed to unlearn data. We therefore propose a family of new loss functions called Subtracted Importance Sampled Scores (SISS) that utilize importance sampling and are the first method to unlearn data with theoretical guarantees. SISS is constructed as a weighted combination between simpler objectives that are responsible for preserving model quality and unlearning the targeted datapoints. When evaluated on CelebA-HQ and MNIST, SISS achieved Pareto optimality along the quality and unlearning strength dimensions. On Stable Diffusion, SISS successfully mitigated memorization on nearly 90% of the prompts we tested. We release our code online.

30. Repurposing in AI: A Distinct Approach or an Extension of Creative Problem Solving?

链接：https://iclr.cc/virtual/2025/poster/31336 abstract： Creativity is defined as the ability to produce novel, useful, and surprising ideas. A sub area of creativity is creative problem solving, the capacity of an agent to discover novel and previously unseen ways to accomplish a task, according to its perspective. While creative problem solving has been extensively studied in AI, the related concept of repurposing - identifying and utilizing existing resources in innovative ways to address different problems from their intended purpose - has received less formal attention. This paper presents a theoretical framework that distinguishes repurposing from creative problem solving by formalizing both approaches in terms of conceptual spaces, resource properties, and goal achievement mechanisms. We demonstrate that while creative problem solving involves expanding the conceptual space through transformation functions, repurposing operates within existing conceptual spaces by leveraging shared properties of available resources. This formalization provides new insights into how these two approaches to

problem-solving differ in their fundamental mechanisms while potentially complementing each other in practical applications.

31. Consistency Models Made Easy

链接：https://iclr.cc/virtual/2025/poster/27780 abstract： Consistency models (CMs) offer faster sampling than traditional diffusion models, but their training is resource-intensive. For example, as of 2024, training a state-of-the-art CM on CIFAR-10 takes one week on 8 GPUs. In this work, we propose an effective scheme for training CMs that largely improves the efficiency of building such models. Specifically, by expressing CM trajectories via a particular differential equation, we argue that diffusion models can be viewed as a special case of CMs. We can thus fine-tune a consistency model starting from a pretrained diffusion model and progressively approximate the full consistency condition to stronger degrees over the training process. Our resulting method, which we term Easy Consistency Tuning (ECT), achieves vastly reduced training times while improving upon the quality of previous methods: for example, ECT achieves a 2-step FID of 2.73 on CIFAR10 within 1 hour on a single A100 GPU, matching Consistency Distillation trained for hundreds of GPU hours. Owing to this computational efficiency, we investigate the scaling laws of CMs under ECT, showing that they obey the classic power law scaling, hinting at their ability to improve efficiency and performance at larger scales. Our code is available.

32. MAD-TD: Model-Augmented Data stabilizes High Update Ratio RL

链接：https://iclr.cc/virtual/2025/poster/30886 abstract： Building deep reinforcement learning (RL) agents that find a good policy with few samples has proven notoriously challenging. To achieve sample efficiency, recent work has explored updating neural networks with large numbers of gradient steps for every new sample. While such high update-to-data (UTD) ratios have shown strong empirical performance, they also introduce instability to the training process. Previous approaches need to rely on periodic neural network parameter resets to address this instability, but restarting the training process is infeasible in many real-world applications and requires tuning the resetting interval. In this paper, we focus on one of the core difficulties of stable training with limited samples: the inability of learned value functions to generalize to unobserved on-policy actions. We mitigate this issue directly by augmenting the off-policy RL training process with a small amount of data generated from a learned world model. Our method, Model-Augmented Data for TD Learning (MAD-TD) uses small amounts of generated data to stabilize high UTD training and achieve competitive performance on the most challenging tasks in the DeepMind control suite. Our experiments further highlight the importance of employing a good model to generate data, MAD-TD's ability to combat value overestimation, and its practical stability gains for continued learning.

33. Fast Uncovering of Protein Sequence Diversity from Structure

链接：https://iclr.cc/virtual/2025/poster/32112 abstract： We present InvMSAFold, an inverse folding method for generating protein sequences optimized for diversity and speed. For a given structure, InvMSAFold generates the parameters of a pairwise probability distribution over the space of sequences, capturing the amino acid covariances observed in Multiple Sequence Alignments (MSA) of homologous proteins. This allows for the efficient generation of highly diverse protein sequences while preserving structural and functional integrity.We demonstrate that this increased diversity in sampled sequences translates into greater variability in biochemical properties, highlighting the exciting potential of our method for applications such as protein design. The orders of magnitude improvement in sampling speed compared to existing methods unlocks new possibilities for high-throughput in virtual screening.

34. Adaptive Data Optimization: Dynamic Sample Selection with Scaling Laws

链接：https://iclr.cc/virtual/2025/poster/29145 abstract： The composition of pretraining data is a key determinant of foundation models' performance, but there is no standard guideline for allocating a limited computational budget across different data sources. Most current approaches either rely on extensive experiments with smaller models or dynamic data adjustments that also require proxy models, both of which significantly increase the workflow complexity and computational overhead. In this paper, we introduce Adaptive Data Optimization (ADO), an algorithm that optimizes data distributions in an online fashion, concurrent with model training. Unlike existing techniques, ADO does not require external knowledge, proxy models, or modifications to the model update. Instead, ADO uses per-domain scaling laws to estimate the learning potential of each domain during training and adjusts the data mixture accordingly, making it more scalable and easier to integrate. Experiments demonstrate that ADO can achieve comparable or better performance than prior methods while maintaining computational efficiency across different computation scales, offering a practical solution for dynamically adjusting data distribution without sacrificing flexibility or increasing costs. Beyond its practical benefits, ADO also provides a new perspective on data collection strategies via scaling laws.

35. Compute-Optimal LLMs Provably Generalize Better with Scale

链接：https://iclr.cc/virtual/2025/poster/29945 abstract： Why do larger language models generalize better? To explore this question, we develop generalization bounds on the pretraining objective of large language models (LLMs) in the compute-optimal regime, as described by the Chinchilla scaling laws. We introduce a novel, fully empirical Freedman-type martingale concentration inequality that tightens existing bounds by accounting for the variance of the loss function. The generalization bound can be broken into three contributions: the number of parameters per token, the loss variance, and the quantization error

at a fixed bitrate. As language models are scaled up, the number of parameters per data point stays constant; however, both the loss variance and the quantization error decrease, implying that larger models should have \emph{smaller} generalization gaps. We examine why larger models tend to be more quantizable from an information theoretic perspective, showing that the rate at which they can integrate new information grows slower than their capacity on the compute optimal frontier. From these findings we produce a scaling law for the generalization gap, showing that our bounds decrease in a predictable way.

36. AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents

链接：https://iclr.cc/virtual/2025/poster/32106 abstract： The robustness of LLMs to jailbreak attacks, where users design prompts to circumvent safety measures and misuse model capabilities, has been studied primarily for LLMs acting as simple chatbots. Meanwhile, LLM agents---which use external tools and can execute multi-stage tasks---may pose a greater risk if misused, but their robustness remains underexplored. To facilitate research on LLM agent misuse, we propose a new benchmark called AgentHarm. The benchmark includes a diverse set of 110 explicitly malicious agent tasks (440 with augmentations), covering 11 harm categories including fraud, cybercrime, and harassment. In addition to measuring whether models refuse harmful agentic requests, scoring well on AgentHarm requires jailbroken agents to maintain their capabilities following an attack to complete a multi-step task. We evaluate a range of leading LLMs, and find (1) leading LLMs are surprisingly complaint with malicious agent requests without jailbreaking, (2) simple universal jailbreak strings can be adapted to effectively jailbreak agents, and (3) these jailbreaks enable coherent and malicious multi-step agent behavior and retain model capabilities. To enable simple and reliable evaluation of attacks and defenses for LLM-based agents, we publicly release AgentHarm at https://huggingface.co/datasets/ai-safety-institute/AgentHarm.

37. Large Language Models Meet Symbolic Provers for Logical Reasoning Evaluation

链接：https://iclr.cc/virtual/2025/poster/30534 abstract： First-order logic (FOL) reasoning, which involves sequential deduction, is pivotal for intelligent systems and serves as a valuable task for evaluating reasoning capabilities, particularly in chain-of-thought (CoT) contexts. Existing benchmarks often rely on extensive human annotation or handcrafted templates, making it difficult to achieve the necessary complexity, scalability, and diversity for robust evaluation. To address these limitations, we propose a novel framework called ProverGen that synergizes the generative strengths of Large Language Models (LLMs) with the rigor and precision of symbolic provers, enabling the creation of a scalable, diverse, and high-quality FOL reasoning dataset, ProverQA. ProverQA is also distinguished by its inclusion of accessible and logically coherent intermediate reasoning steps for each problem. Our evaluation shows that state-of-the-art LLMs struggle to solve ProverQA problems, even with CoT prompting, highlighting the dataset's challenging nature. We also finetune Llama3.1-8B-Instruct on a separate training set generated by our framework.The finetuned model demonstrates consistent improvements on both in-distribution and out-of-distribution test sets, suggesting the value of our proposed data generation framework. Code available at: \url{https://github.com/opendatalab/ProverGen}

38. Learning mirror maps in policy mirror descent

链接：https://iclr.cc/virtual/2025/poster/28430 abstract： Policy Mirror Descent (PMD) is a popular framework in reinforcement learning, serving as a unifying perspective that encompasses numerous algorithms. These algorithms are derived through the selection of a mirror map and enjoy finite-time convergence guarantees. Despite its popularity, the exploration of PMD's full potential is limited, with the majority of research focusing on a particular mirror map---namely, the negative entropy---which gives rise to the renowned Natural Policy Gradient (NPG) method. It remains uncertain from existing theoretical studies whether the choice of mirror map significantly influences PMD's efficacy. In our work, we conduct empirical investigations to show that the conventional mirror map choice (NPG) often yields less-than-optimal outcomes across several standard benchmark environments. Using evolutionary strategies, we identify more efficient mirror maps that enhance the performance of PMD. We first focus on a tabular environment, i.e.\ Grid-World, where we relate existing theoretical bounds with the performance of PMD for a few standard mirror maps and the learned one. We then show that it is possible to learn a mirror map that outperforms the negative entropy in more complex environments, such as the MinAtar suite. Additionally, we demonstrate that the learned mirror maps generalize effectively to different tasks by testing each map across various other environments.

39. DiTTo-TTS: Diffusion Transformers for Scalable Text-to-Speech without Domain-Specific Factors

链接：https://iclr.cc/virtual/2025/poster/32067 abstract： Large-scale latent diffusion models (LDMs) excel in content generation across various modalities, but their reliance on phonemes and durations in text-to-speech (TTS) limits scalability and access from other fields. While recent studies show potential in removing these domain-specific factors, performance remains suboptimal. In this work, we introduce DiTTo-TTS, a Diffusion Transformer (DiT)-based TTS model, to investigate whether LDM-based TTS can achieve state-of-the-art performance without domain-specific factors. Through rigorous analysis and empirical exploration, we find that (1) DiT with minimal modifications outperforms U-Net, (2) variable-length modeling with a speech length predictor significantly improves results over fixed-length approaches, and (3) conditions like semantic alignment in speech latent representations are key to further enhancement. By scaling our training data to 82K hours and the model size to 790M parameters, we achieve superior or comparable zero-shot performance to state-of-the-art TTS models in naturalness, intelligibility, and speaker similarity, all without relying on domain-specific factors. Speech samples are available at https://ditto-

tts.github.io.

40. CodeMMLU: A Multi-Task Benchmark for Assessing Code Understanding & Reasoning Capabilities of CodeLLMs

链接：https://iclr.cc/virtual/2025/poster/30509 abstract： Recent advances in Code Large Language Models (CodeLLMs) have primarily focused on open-ended code generation, often overlooking the crucial aspect of code understanding & reasoning. To bridge this gap, we introduce CodeMMLU, a comprehensive multiple-choice benchmark designed to evaluate the depth of software and code comprehension in LLMs. CodeMMLU includes nearly 20,000 questions spanning diverse domains, including code analysis, defect detection, and software engineering principles across multiple programming languages. Unlike traditional benchmarks that emphasize code generation, CodeMMLU assesses a model's ability to reason about programs across a wide-range of tasks such as code repair, execution reasoning, and fill-in-the-blank challenges. Our extensive evaluation reveals that even state-of-the-art models struggle with CodeMMLU, highlighting significant gaps in comprehension beyond generation. By emphasizing the essential connection between code understanding and effective AI-assisted development, CodeMMLU provides a critical resource for advancing more reliable and capable coding assistants.

41. Bridging the Data Provenance Gap Across Text, Speech, and Video

链接：https://iclr.cc/virtual/2025/poster/30303 abstract： Progress in AI is driven largely by the scale and quality of training data. Despite this, there is a deficit of empirical analysis examining the attributes of well-established datasets beyond text. In this work we conduct the largest and first-of-its-kind longitudinal audit across modalities --- popular text, speech, and video datasets --- from their detailed sourcing trends and use restrictions to their geographical and linguistic representation. Our manual analysis covers nearly 4000 public datasets between 1990-2024, spanning 608 languages, 798 sources, 659 organizations, and 67 countries. We find that multimodal machine learning applications have overwhelmingly turned to web-crawled, synthetic, and social media platforms, such as YouTube, for their training sets, eclipsing all other sources since 2019. Secondly, tracing the chain of dataset derivations we find that while less than 33% of datasets are restrictively licensed, over 80% of the source content in widely-used text, speech, and video datasets, carry non-commercial restrictions. Finally, counter to the rising number of languages and geographies represented in public AI training datasets, our audit demonstrates measures of relative geographical and multilingual representation have failed to significantly improve their coverage since 2013. We believe the breadth of our audit enables us to empirically examine trends in data sourcing, restrictions, and Western-centricity at an ecosystem-level, and that visibility into these questions are essential to progress in responsible AI. As a contribution to ongoing improvements in dataset transparency and responsible use, we release our entire multimodal audit, allowing practitioners to trace data provenance across text, speech, and video.

42. Decoupled Subgraph Federated Learning

链接：https://iclr.cc/virtual/2025/poster/27937 abstract： We address the challenge of federated learning on graph-structured data distributed across multiple clients. Specifically, we focus on the prevalent scenario of interconnected subgraphs, where inter-connections between different clients play a critical role. We present a novel framework for this scenario, named FedStruct, that harnesses deep structural dependencies. To uphold privacy, unlike existing methods, FedStruct eliminates the necessity of sharing or generating sensitive node features or embeddings among clients. Instead, it leverages explicit global graph structure information to capture inter-node dependencies. We validate the effectiveness of FedStruct through experimental results conducted on six datasets for semi-supervised node classification, showcasing performance close to the centralized approach across various scenarios, including different data partitioning methods, varying levels of label availability, and number of clients.

43. Improved Diffusion-based Generative Model with Better Adversarial Robustness

链接：https://iclr.cc/virtual/2025/poster/31218 abstract： Diffusion Probabilistic Models (DPMs) have achieved significant success in generative tasks. However, their training and sampling processes suffer from the issue of distribution mismatch. During the denoising process, the input data distributions differ between the training and inference stages, potentially leading to inaccurate data generation. To obviate this, we analyze the training objective of DPMs and theoretically demonstrate that this mismatch can be alleviated through Distributionally Robust Optimization (DRO), which is equivalent to performing robustness-driven Adversarial Training (AT) on DPMs. Furthermore, for the recently proposed Consistency Model (CM), which distills the inference process of the DPM, we prove that its training objective also encounters the mismatch issue. Fortunately, this issue can be mitigated by AT as well. Based on these insights, we propose to conduct efficient AT on both DPM and CM. Finally, extensive empirical studies validate the effectiveness of AT in diffusion-based models. The code is available at https://github.com/kugwzk/AT_Diff.

44. Manifold Learning by Mixture Models of VAEs for Inverse Problems

链接：https://iclr.cc/virtual/2025/poster/31386 abstract： Representing a manifold of very high-dimensional data with generative models has been shown to be computationally efficient in practice. However, this requires that the data manifold admits a global parameterization. In order to represent manifolds of arbitrary topology, we propose to learn a mixture model of variational

autoencoders. Here, every encoder-decoder pair represents one chart of a manifold. We propose a loss function for maximum likelihood estimation of the model weights and choose an architecture that provides us the analytical expression of the charts and of their inverses. Once the manifold is learned, we use it for solving inverse problems by minimizing a data fidelity term restricted to the learned manifold. To solve the arising minimization problem we propose a Riemannian gradient descent algorithm on the learned manifold. We demonstrate the performance of our method for low-dimensional toy examples as well as for deblurring and electrical impedance tomography on certain image manifolds.

45. Generative Classifiers Avoid Shortcut Solutions

链接：https://iclr.cc/virtual/2025/poster/28371 abstract： Discriminative approaches to classification often learn shortcuts that hold in-distribution but fail even under minor distribution shift. This failure mode stems from an overreliance on features that are spuriously correlated with the label. We show that generative classifiers, which use class-conditional generative models, can avoid this issue by modeling all features, both core and spurious, instead of mainly spurious ones. These generative classifiers are simple to train, avoiding the need for specialized augmentations, strong regularization, extra hyperparameters, or knowledge of the specific spurious correlations to avoid. We find that diffusion-based and autoregressive generative classifiers achieve state-of-the-art performance on five standard image and text distribution shift benchmarks and reduce the impact of spurious correlations in realistic applications, such as medical or satellite datasets. Finally, we carefully analyze a Gaussian toy setting to understand the inductive biases of generative classifiers, as well as the data properties that determine when generative classifiers outperform discriminative ones.

46. AgentTrek: Agent Trajectory Synthesis via Guiding Replay with Web Tutorials

链接：https://iclr.cc/virtual/2025/poster/30416 abstract： Graphical User Interface (GUI) agents hold great potential for automating complex tasks across diverse digital environments, from web applications to desktop software. However, the development of such agents is hindered by the lack of high-quality, multi-step trajectory data required for effective training. Existing approaches rely on expensive and labor-intensive human annotation, making them unsustainable at scale. To address this challenge, we propose AgentTrek, a scalable data synthesis pipeline that generates high-quality web agent trajectories by leveraging web tutorials. Our method automatically gathers tutorial-like texts from the internet, transforms them into task goals with step-by-step instructions, and employs a visual-language model (VLM) agent to simulate their execution in a real digital environment. A VLM-based evaluator ensures the correctness of the generated trajectories. We demonstrate that training GUI agents with these synthesized trajectories significantly improves their grounding and planning performance over the current models. Moreover, our approach is more cost-efficient compared to traditional human annotation methods. This work underscores the potential of guided replay with web tutorials as a viable strategy for large-scale GUI agent training, paving the way for more capable and autonomous digital agents.

47. Everything, Everywhere, All at Once: Is Mechanistic Interpretability Identifiable?

链接：https://iclr.cc/virtual/2025/poster/30956 abstract： As AI systems are increasingly deployed in high-stakes applications, ensuring their interpretability is essential. Mechanistic Interpretability (MI) aims to reverse-engineer neural networks by extracting human-understandable algorithms embedded within their structures to explain their behavior. This work systematically examines a fundamental question: for a fixed behavior to explain, and under the criteria that MI sets for itself, are we guaranteed a unique explanation? Drawing an analogy with the concept of identifiability in statistics, which ensures the uniqueness of parameters inferred from data under specific modeling assumptions, we speak about the identifiability of explanations produced by MI.We identify two broad strategies to produce MI explanations: (i) "where-then-what", which first identifies a subset of the network (a circuit) that replicates the model's behavior before deriving its interpretation, and (ii) "what-then-where", which begins with candidate explanatory algorithms and searches in the activation subspaces of the neural model where the candidate algorithm may be implemented, relying on notions of causal alignment between the states of the candidate algorithm and the neural network. We systematically test the identifiability of both strategies using simple tasks (learning Boolean functions) and multi-layer perceptrons small enough to allow a complete enumeration of candidate explanations. Our experiments reveal overwhelming evidence of non-identifiability in all cases: multiple circuits can replicate model behavior, multiple interpretations can exist for a circuit, several algorithms can be causally aligned with the neural network, and a single algorithm can be causally aligned with different subspaces of the network.We discuss whether the unicity intuition is necessary. One could adopt a pragmatic stance, requiring explanations only to meet predictive and/or manipulability standards. However, if unicity is considered essential, e.g., to provide a sense of understanding, we also discuss less permissive criteria. Finally, we also refer to the inner interpretability framework that demands explanations to be validated by multiple complementary criteria. This work aims to contribute constructively to the ongoing effort to formalize what we expect from explanations in AI.

48. Masked Diffusion Models are Secretly Time-Agnostic Masked Models and Exploit Inaccurate Categorical Sampling

链接：https://iclr.cc/virtual/2025/poster/30511 abstract： Masked diffusion models (MDMs) have emerged as a popular research topic for generative modeling of discrete data, thanks to their superior performance over other discrete diffusion

models, and are rivaling the auto-regressive models (ARMs) for language modeling tasks. The recent effort in simplifying the masked diffusion framework further leads to alignment with continuous-space diffusion models and more principled training and sampling recipes. In this paper, however, we reveal that both training and sampling of MDMs are theoretically free from the time variable, arguably the key signature of diffusion models, and are instead equivalent to masked models. The connection on the sampling aspect is drawn by our proposed first-hitting sampler (FHS). Specifically, we show that the FHS is theoretically equivalent to MDMs' original generation process while significantly alleviating the time-consuming categorical sampling and achieving a 20$\times$ speedup. In addition, our investigation raises doubts about whether MDMs can truly beat ARMs in text generation. We identify, for the first time, an underlying numerical issue, even with the commonly used 32-bit floating-point precision, which results in inaccurate categorical sampling. We show that it lowers the effective temperature both theoretically and empirically, and the resulting decrease in token diversity makes previous evaluations, which assess the generation quality solely through the incomplete generative perplexity metric, somewhat unfair.

49. Diffusion Bridge Implicit Models

链接：https://iclr.cc/virtual/2025/poster/28911 abstract： Denoising diffusion bridge models (DDBMs) are a powerful variant of diffusion models for interpolating between two arbitrary paired distributions given as endpoints. Despite their promising performance in tasks like image translation, DDBMs require a computationally intensive sampling process that involves the simulation of a (stochastic) differential equation through hundreds of network evaluations. In this work, we take the first step in fast sampling of DDBMs without extra training, motivated by the well-established recipes in diffusion models. We generalize DDBMs via a class of non-Markovian diffusion bridges defined on the discretized timesteps concerning sampling, which share the same marginal distributions and training objectives, give rise to generative processes ranging from stochastic to deterministic, and result in diffusion bridge implicit models (DBIMs). DBIMs are not only up to 25$\times$ faster than the vanilla sampler of DDBMs but also induce a novel, simple, and insightful form of ordinary differential equation (ODE) which inspires high-order numerical solvers. Moreover, DBIMs maintain the generation diversity in a distinguished way, by using a booting noise in the initial sampling step, which enables faithful encoding, reconstruction, and semantic interpolation in image translation tasks. Code is available at \url{https://github.com/thu-ml/DiffusionBridge}.

50. LiFT: Learning to Fine-Tune via Bayesian Parameter Efficient Meta Fine-Tuning

链接：https://iclr.cc/virtual/2025/poster/30801 abstract： We tackle the problem of parameter-efficient fine-tuning (PEFT) of a pre-trained large deep model on many different but related tasks. Instead of the simple but strong baseline strategy of task-wise independent fine-tuning, we aim to meta-learn the core shared information that can be used for unseen test tasks to improve the prediction performance further. That is, we propose a method for {\em learning-to-fine-tune} (LiFT). LiFT introduces a novel hierarchical Bayesian model that can be superior to both existing general meta learning algorithms like MAML and recent LoRA zoo mixing approaches such as LoRA-Retriever and model-based clustering. In our Bayesian model, the parameters of the task-specific LoRA modules are regarded as random variables where these task-wise LoRA modules are governed/regularized by higher-level latent random variables, which represents the prior of the LoRA modules that capture the shared information across all training tasks. To make the posterior inference feasible, we propose a novel SGLD-Gibbs sampling algorithm that is computationally efficient. To represent the posterior samples from the SGLD-Gibbs, we propose an online EM algorithm that maintains a Gaussian mixture representation for the posterior in an online manner in the course of iterative posterior sampling. We demonstrate the effectiveness of LiFT on NLP and vision multi-task meta learning benchmarks.

51. Beyond Worst-Case Dimensionality Reduction for Sparse Vectors

链接：https://iclr.cc/virtual/2025/poster/28947 abstract： We study beyond worst-case dimensionality reduction for s-sparse vectors (vectors with at most s non-zero coordinates). Our work is divided into two parts, each focusing on a different facet of beyond worst-case analysis:\noindent (a) We first consider average-case guarantees for embedding s-sparse vectors. Here, a well-known folklore upper bound based on the birthday-paradox states: For any collection X of s-sparse vectors in $\mathbb{R}^d$, there exists a linear map $A: \mathbb{R}^d \rightarrow \mathbb{R}^{O(s^2)}$ which \emph{exactly} preserves the norm of 99% of the vectors in X in any ℓ_p norm (as opposed to the usual setting where guarantees hold for all vectors). We provide novel lower bounds showing that this is indeed optimal in many settings. Specifically, any oblivious linear map satisfying similar average-case guarantees must map to $\Omega(s^2)$ dimensions. The same lower bound also holds for a wider class of sufficiently smooth maps, including `encoder-decoder schemes', where we compare the norm of the original vector to that of a smooth function of the embedding. These lower bounds reveal a surprising separation result for smooth embeddings of sparse vectors, as an upper bound of $O(s \log(d))$ is possible if we instead use arbitrary functions, e.g., via compressed sensing algorithms. (b) Given these lower bounds, we specialize to sparse \emph{non-negative} vectors to hopes of improved upper bounds. For a dataset X of non-negative s-sparse vectors and any $p \ge 1$, we can non-linearly embed X to $O(s\log(|X|s)/\varepsilon^2)$ dimensions while preserving all pairwise distances in ℓ_p norm up to $1\pm \varepsilon$, with no dependence on p. Surprisingly, the non-negativity assumption enables much smaller embeddings than arbitrary sparse vectors, where the best known bound suffers an exponential $(\log |X|)^{O(p)}$ dependence. Our map also guarantees \emph{exact} dimensionality reduction for the ℓ_{∞} norm by embedding X into $O(s\log |X|)$ dimensions, which is tight. We further give separation results showing that both the non-linearity of f and the non-negativity of X are necessary, and provide downstream algorithmic improvements using our embedding.

52. Elucidating the Preconditioning in Consistency Distillation

链接：https://iclr.cc/virtual/2025/poster/30971 abstract： Consistency distillation is a prevalent way for accelerating diffusion models adopted in consistency (trajectory) models, in which a student model is trained to traverse backward on the probability flow (PF) ordinary differential equation (ODE) trajectory determined by the teacher model. Preconditioning is a vital technique for stabilizing consistency distillation, by linear combining the input data and the network output with pre-defined coefficients as the consistency function. It imposes the boundary condition of consistency functions without restricting the form and expressiveness of the neural network. However, previous preconditionings are hand-crafted and may be suboptimal choices. In this work, we offer the first theoretical insights into the preconditioning in consistency distillation, by elucidating its design criteria and the connection to the teacher ODE trajectory. Based on these analyses, we further propose a principled way dubbed \textit{Analytic-Precond} to analytically optimize the preconditioning according to the consistency gap (defined as the gap between the teacher denoiser and the optimal student denoiser) on a generalized teacher ODE. We demonstrate that Analytic-Precond can facilitate the learning of trajectory jumpers, enhance the alignment of the student trajectory with the teacher's, and achieve $2\times$ to $3\times$ training acceleration of consistency trajectory models in multi-step generation across various datasets.

53. Improving Data Efficiency via Curating LLM-Driven Rating Systems

链接：https://iclr.cc/virtual/2025/poster/30465 abstract： Instruction tuning is critical for adapting large language models (LLMs) to downstream tasks, and recent studies have demonstrated that small amounts of human-curated data can outperform larger datasets, challenging traditional data scaling laws. While LLM-based data quality rating systems offer a cost-effective alternative to human annotation, they often suffer from inaccuracies and biases, even in powerful models like GPT-4. In this work, we introduce DS^2, a **D**iversity-aware **S**core curation method for **D**ata **S**election. By systematically modeling error patterns through a score transition matrix, DS^2 corrects LLM-based scores and promotes diversity in the selected data samples. Our approach shows that a curated subset (just 3.3\% of the original dataset) outperforms full-scale datasets (300k samples) across various machine-alignment benchmarks, and matches or surpasses human-aligned datasets such as LIMA with the same sample size (1k samples). These findings challenge conventional data scaling assumptions, highlighting that redundant, low-quality samples can degrade performance and reaffirming that ``more can be less".

54. Convergence of Score-Based Discrete Diffusion Models: A Discrete-Time Analysis

链接：https://iclr.cc/virtual/2025/poster/28280 abstract： Diffusion models have achieved great success in generating high-dimensional samples across various applications. While the theoretical guarantees for continuous-state diffusion models have been extensively studied, the convergence analysis of the discrete-state counterparts remains under-explored. In this paper, we study the theoretical aspects of score-based discrete diffusion models under the Continuous Time Markov Chain (CTMC) framework. We introduce a discrete-time sampling algorithm in the general state space $[S]^d$ that utilizes score estimators at predefined time points. We derive convergence bounds for the Kullback-Leibler (KL) divergence and total variation (TV) distance between the generated sample distribution and the data distribution, considering both scenarios with and without early stopping under reasonable assumptions. Notably, our KL divergence bounds are nearly linear in the dimension d, aligning with state-of-the-art results for diffusion models. Our convergence analysis employs a Girsanov-based method and establishes key properties of the discrete score function, which are essential for characterizing the discrete-time sampling process.

55. Retrieval Head Mechanistically Explains Long-Context Factuality

链接：https://iclr.cc/virtual/2025/poster/30373 abstract： Despite the recent progress in long-context language models, it remains elusive how transformer-based models exhibit the capability to retrieve relevant information from arbitrary locations within the long context. This paper aims to address this question. Our systematic investigation across a wide spectrum of models reveals that a special type of attention heads are largely responsible for retrieving information, which we dub retrieval heads. We identify intriguing properties of retrieval heads:(1) universal: all the explored models with long-context capability have a set of retrieval heads; (2) sparse: only a small portion (less than 5\%) of the attention heads are retrieval. (3) intrinsic: retrieval heads already exist in models pretrained with short context. When extending the context length by continual pretraining, it is still the same set of heads that perform information retrieval. (4) dynamically activated: take Llama-2 7B for example, 12 retrieval heads always attend to the required information no matter how the context is changed. The rest of the retrieval heads are activated in different contexts. (5) causal: completely pruning retrieval heads leads to failure in retrieving relevant information and results in hallucination, while pruning random non-retrieval heads does not affect the model's retrieval ability. We further show that retrieval heads strongly influence chain-of-thought (CoT) reasoning, where the model needs to frequently refer back the question and previously-generated context. Conversely, tasks where the model directly generates the answer using its intrinsic knowledge are less impacted by masking out retrieval heads. These observations collectively explain which internal part of the model seeks information from the input tokens. We believe our insights will foster future research on reducing hallucination, improving reasoning, and compressing the KV cache.

56. Do LLMs estimate uncertainty well in instruction-following?

链接：https://iclr.cc/virtual/2025/poster/30179 abstract： Large language models (LLMs) could be valuable personal AI agents across various domains, provided they can precisely follow user instructions. However, recent studies have shown significant limitations in LLMs' instruction-following capabilities, raising concerns about their reliability in high-stakes applications.

Accurately estimating LLMs' uncertainty in adhering to instructions is critical to mitigating deployment risks. We present, to our knowledge, the first systematic evaluation of the uncertainty estimation abilities of LLMs in the context of instruction-following. Our study identifies key challenges with existing instruction-following benchmarks, where multiple factors are entangled with uncertainty stems from instruction-following, complicating the isolation and comparison across methods and models.To address these issues, we introduce a controlled evaluation setup with two benchmark versions of data, enabling a comprehensive comparison of uncertainty estimation methods under various conditions.Our findings show that existing uncertainty methods struggle, particularly when models make subtle errors in instruction following. While internal model states provide some improvement, they remain inadequate in more complex scenarios. The insights from our controlled evaluation setups provide a crucial understanding of LLMs' limitations and potential for uncertainty estimation in instruction-following tasks, paving the way for more trustworthy AI agents.

57. Nonconvex Stochastic Optimization under Heavy-Tailed Noises: Optimal Convergence without Gradient Clipping

链接：https://iclr.cc/virtual/2025/poster/29883 abstract： Recently, the study of heavy-tailed noises in first-order nonconvex stochastic optimization has gotten a lot of attention since it was recognized as a more realistic condition as suggested by many empirical observations. Specifically, the stochastic noise (the difference between the stochastic and true gradient) is considered to have only a finite $\mathfrak{p}$-th moment where $\mathfrak{p}\in\left(1,2\right]$ instead of assuming it always satisfies the classical finite variance assumption. To deal with this more challenging setting, people have proposed different algorithms and proved them to converge at an optimal $\mathcal{O}(T^{\frac{1-\mathfrak{p}}{3\mathfrak{p}-2}})$ rate for smooth objectives after T iterations. Notably, all these new-designed algorithms are based on the same technique – gradient clipping. Naturally, one may want to know whether the clipping method is a necessary ingredient and the only way to guarantee convergence under heavy-tailed noises. In this work, by revisiting the existing Batched Normalized Stochastic Gradient Descent with Momentum (Batched NSGDM) algorithm, we provide the first convergence result under heavy-tailed noises but without gradient clipping. Concretely, we prove that Batched NSGDM can achieve the optimal $\mathcal{O}(T^{\frac{1-\mathfrak{p}}{3\mathfrak{p}-2}})$ rate even under the relaxed smooth condition. More interestingly, we also establish the first $\mathcal{O}(T^{\frac{1-\mathfrak{p}}{2\mathfrak{p}}})$ convergence rate in the case where the tail index $\mathfrak{p}$ is unknown in advance, which is arguably the common scenario in practice.

58. Large (Vision) Language Models are Unsupervised In-Context Learners

链接：https://iclr.cc/virtual/2025/poster/28335 abstract： Recent advances in large language and vision-language models have enabled zero-shot inference, allowing models to solve new tasks without task-specific training. Various adaptation techniques such as prompt engineering, In-Context Learning (ICL), and supervised fine-tuning can further enhance the model's performance on a downstream task, but they require substantial manual effort to construct effective prompts or labeled examples. In this work, we introduce a joint inference framework for fully unsupervised adaptation, eliminating the need for manual prompt engineering and labeled examples. Unlike zero-shot inference, which makes independent predictions, the joint inference makes predictions simultaneously for all inputs in a given task. Since direct joint inference involves computationally expensive optimization, we develop efficient approximation techniques, leading to two unsupervised adaptation methods: unsupervised fine-tuning and unsupervised ICL. We demonstrate the effectiveness of our methods across diverse tasks and models, including language-only Llama-3.1 on natural language processing tasks, reasoning-oriented Qwen2.5-Math on grade school math problems, vision-language OpenFlamingo on vision tasks, and the API-only access GPT-4o model on massive multi-discipline tasks. Our experiments demonstrate substantial improvements over the standard zero-shot approach, including 39% absolute improvement on the challenging GSM8K math reasoning dataset. Remarkably, despite being fully unsupervised, our framework often performs on par with supervised approaches that rely on ground truth labels.

59. ThunderKittens: Simple, Fast, and $\textit{Adorable}$ Kernels

链接：https://iclr.cc/virtual/2025/poster/31243 abstract： The challenge of mapping AI architectures to GPU hardware is creating a critical bottleneck in AI progress. Despite substantial efforts, hand-written custom kernels fail to meet their theoretical performance thresholds, even on well-established operations like linear attention. The diverse capabilities of GPUs suggests we might we need a wide variety of techniques to achieve high performance. However, our work explores if a small number of key abstractions can drastically simplify the process. We present ThunderKittens (TK), a framework for writing performant AI kernels while remaining easy to use. Our abstractions map to the three levels of the GPU hierarchy: (1) at the warp-level, we provide 16x16 matrix tiles as basic data structures and PyTorch-like operations, (2) at the thread-block level, we provide templates for asynchronously overlapping operations, and (3) at the grid-level, TK helps hide block launch, tear-down, and memory costs. We show the value of TK by providing simple & diverse kernels that match or outperform prior art. We match CuBLAS and FlashAttention-3 on GEMM and attention inference performance and outperform the strongest baselines by $10-40$\% on attention backwards, $8\times$ on state space models, and $14\times$ on linear attention.

60. Achieving Dimension-Free Communication in Federated Learning via Zeroth-Order Optimization

链接：https://iclr.cc/virtual/2025/poster/28334 abstract： Federated Learning (FL) offers a promising framework for collaborative and privacy-preserving machine learning across distributed data sources. However, the substantial communication

costs associated with FL significantly challenge its efficiency. Specifically, in each communication round, the communication costs scale linearly with the model's dimension, which presents a formidable obstacle, especially in large model scenarios. Despite various communication-efficient strategies, the intrinsic dimension-dependent communication cost remains a major bottleneck for current FL implementations.This paper proposes a novel dimension-free communication algorithm - DeComFL, which leverages the zeroth-order optimization techniques and reduces the communication cost from $\mathcal{O}(d)$ to $\mathcal{O}(1)$ by transmitting only a constant number of scalar values between clients and the server in each round, regardless of the dimension d of the model parameters.Theoretically, in non-convex functions, we prove that our algorithm achieves state-of-the-art rates, which show a linear speedup of the number of clients and local steps under standard assumptions. With additional low effective rank assumption, we can further show that the convergence rate is independent of the model dimension d as well.Empirical evaluations, encompassing both classic deep learning training and large language model fine-tuning, demonstrate significant reductions in communication overhead. Notably, DeComFL achieves this by transmitting only around 1MB of data in total between the server and a client to fine-tune a model with billions of parameters. The code is available at https://github.com/ZidongLiu/DeComFL.

61. Chain-of-Thought Provably Enables Learning the (Otherwise) Unlearnable

链接：https://iclr.cc/virtual/2025/poster/29898 abstract： Modern language models have demonstrated remarkable reasoning capabilities by using chain-of-thought (CoT). One hypothesis about the inner workings of CoT is that it breaks down originally complex tasks into smaller subtasks that are more amenable to learning. We formalize this notion by showing possibility and impossibility results of learning from in-context demonstrations with and without CoT. In particular, with CoT, we examine a family of learning algorithms that learn a task step-by-step, capable of composing simpler functions from individual reasoning steps to form an overall complex function. This process reduces the difficulty of learning a task to that of the hardest reasoning step in the chain. Moreover, we prove Transformers can express this algorithm and thus they can efficiently in-context learn arbitrary tasks as long as these tasks can be decomposed into a finite number of subtasks, each of which are efficiently learnable. In contrast, without CoT, we demonstrate that there exist tasks that are inherently unlearnable by the same algorithm. Overall, our results suggest several provably effective ways for decomposing target problems to instantiate CoT. Empirically, we demonstrate our proposed CoT construction significantly enhances the reasoning capabilities of real-world LLMs in solving challenging arithmetic reasoning tasks, including learning polynomials and Boolean formulas.

62. Towards Learning High-Precision Least Squares Algorithms with Sequence Models

链接：https://iclr.cc/virtual/2025/poster/28098 abstract： This paper investigates whether sequence models can learn to perform numerical algorithms, e.g. gradient descent, on the fundamental problem of least squares. Our goal is to inherit two properties of standard algorithms from numerical analysis: (1) machine precision, i.e. we want to obtain solutions that are accurate to near floating point error, and (2) numerical generality, i.e. we want them to apply broadly across problem instances. We find that prior approaches using Transformers fail to meet these criteria, and identify limitations present in existing architectures and training procedures. First, we show that softmax Transformers struggle to perform high-precision multiplications, which prevents them from precisely learning numerical algorithms. Second, we identify an alternate class of architectures, comprised entirely of polynomials, that can efficiently represent high-precision gradient descent iterates. Finally, we investigate precision bottlenecks during training and address them via a high-precision training recipe that reduces stochastic gradient noise. Our recipe enables us to train two polynomial architectures, gated convolutions and linear attention, to perform gradient descent iterates on least squares problems. For the first time, we demonstrate the ability to train to near machine precision. Applied iteratively, our models obtain $100,000\times$ lower MSE than standard Transformers trained end-to-end and they incur a $10,000\times$ smaller generalization gap on out-of-distribution problems. We make progress towards end-to-end learning of numerical algorithms for least squares.

63. Understanding Warmup-Stable-Decay Learning Rates: A River Valley Loss Landscape View

链接：https://iclr.cc/virtual/2025/poster/28484 abstract： Training language models currently requires pre-determining a fixed compute budget because the typical cosine learning rate schedule depends on the total number of steps. In contrast, the Warmup-Stable-Decay (WSD) schedule uses a constant learning rate to produce a main branch of iterates that can in principle continue indefinitely without a pre-specified compute budget. Then, given any compute budget, one can branch out from the main branch at a proper time with a rapidly decaying learning rate to produce a strong model. Empirically, WSD generates an intriguing, non-traditional loss curve: the loss remains elevated during the stable phase but sharply declines during the decay phase. Towards explaining this phenomenon, we conjecture that pretraining loss exhibits a river valley landscape, which resembles a deep valley with a river at its bottom. Under this assumption, we show that during the stable phase, the iterate undergoes large oscillations due to the high learning rate, yet it progresses swiftly along the river. During the decay phase, the rapidly dropping learning rate minimizes the iterate’s oscillations, moving it closer to the river and revealing true optimization progress. Therefore, the sustained high learning rate phase and fast decaying phase are responsible for progress in the river and the mountain directions, respectively, and are both critical. Our analysis predicts phenomenons consistent with empirical observations and shows that this landscape can naturally emerge from pretraining on a simple bi-gram dataset. Inspired by the theory, we introduce WSD-S, a variant of WSD that reuses previous checkpoints’ decay phases and keeps only one main

branch, where we resume from a decayed checkpoint. WSD-S empirically outperforms WSD and Cyclic-Cosine in obtaining multiple pretrained language model checkpoints across various compute budgets in a single run for parameters scaling from 0.1B to 1.2B.

64. nGPT: Normalized Transformer with Representation Learning on the Hypersphere

链接：https://iclr.cc/virtual/2025/poster/28110 abstract： We propose a novel neural network architecture, the normalized Transformer (nGPT) with representation learning on the hypersphere. In nGPT, all vectors forming the embeddings, MLP, attention matrices and hidden states are unit norm normalized. The input stream of tokens travels on the surface of a hypersphere, with each layer contributing a displacement towards the target output predictions. These displacements are defined by the MLP and attention blocks, whose vector components also reside on the same hypersphere. Experiments show that nGPT learns much faster, reducing the number of training steps required to achieve the same accuracy by a factor of 4 to 20, depending on the sequence length.

65. A Coefficient Makes SVRG Effective

链接：https://iclr.cc/virtual/2025/poster/28009 abstract： Stochastic Variance Reduced Gradient (SVRG), introduced by Johnson & Zhang (2013), is a theoretically compelling optimization method. However, as Defazio & Bottou (2019) highlight, its effectiveness in deep learning is yet to be proven. In this work, we demonstrate the potential of SVRG in optimizing real-world neural networks. Our empirical analysis finds that, for deeper neural networks, the strength of the variance reduction term in SVRG should be smaller and decrease as training progresses. Inspired by this, we introduce a multiplicative coefficient α to control the strength and adjust it through a linear decay schedule. We name our method α-SVRG. Our results show α-SVRG better optimizes models, consistently reducing training loss compared to the baseline and standard SVRG across various model architectures and multiple image classification datasets. We hope our findings encourage further exploration into variance reduction techniques in deep learning. Code is available at github.com/davidyyd/alpha-SVRG.

66. Homomorphism Counts as Structural Encodings for Graph Learning

链接：https://iclr.cc/virtual/2025/poster/28259 abstract： Graph Transformers are popular neural networks that extend the well-known Transformer architecture to the graph domain. These architectures operate by applying self-attention on graph nodes and incorporating graph structure through the use of positional encodings (e.g., Laplacian positional encoding) or structural encodings (e.g., random-walk structural encoding). The quality of such encodings is critical, since they provide the necessary \emph{graph inductive biases} to condition the model on graph structure. In this work, we propose \emph{motif structural encoding} (MoSE) as a flexible and powerful structural encoding framework based on counting graph homomorphisms. Theoretically, we compare the expressive power of MoSE to random-walk structural encoding and relate both encodings to the expressive power of standard message passing neural networks. Empirically, we observe that MoSE outperforms other well-known positional and structural encodings across a range of architectures, and it achieves state-of-the-art performance on a widely studied molecular property prediction dataset.

67. Reasoning with Latent Thoughts: On the Power of Looped Transformers

链接：https://iclr.cc/virtual/2025/poster/28971 abstract： Large language models have shown remarkable reasoning abilities and scaling laws suggest that large parameter count, especially along the depth axis, is the primary driver. In this work, we make a stronger claim --- many reasoning problems require a large depth but not necessarily many parameters. This unlocks a novel application of looped models for reasoning. Firstly, we show that for many synthetic reasoning problems like addition, p-hop induction, and math problems, a k-layer transformer looped L times nearly matches the performance of a kL-layer non-looped model, and is significantly better than a k-layer model. This is further corroborated by theoretical results showing that many such reasoning problems can be solved via iterative algorithms, and thus, can be solved effectively using looped models with nearly optimal depth. Perhaps surprisingly, these benefits also translate to practical settings of language modeling --- on many downstream reasoning tasks, a language model with k-layers looped L times can be competitive to, if not better than, a kL-layer language model. In fact, our empirical analysis reveals an intriguing phenomenon: looped and non-looped models exhibit scaling behavior that depends on their effective depth, akin to the inference-time scaling of chain-of-thought (CoT) reasoning. We further elucidate the connection to CoT reasoning by proving that looped models implicitly generate latent thoughts and can simulate T steps of CoT with T loops. Inspired by these findings, we also present an interesting dichotomy between reasoning and memorization, and design a looping-based regularization that is effective on both fronts.

68. PhysPDE: Rethinking PDE Discovery and a Physical HYpothesis Selection Benchmark

链接：https://iclr.cc/virtual/2025/poster/30306 abstract： Despite extensive research, recovering PDE expressions from experimental observations often involves symbolic regression. This method generally lacks the incorporation of meaningful physical insights, resulting in outcomes lacking clear physical interpretations. Recognizing that the primary interest of Machine Learning for Science (ML4Sci) often lies in understanding the underlying physical mechanisms or even discovering new physical

laws rather than simply obtaining mathematical expressions, this paper introduces a novel ML4Sci task paradigm. This paradigm focuses on interpreting experimental data within the framework of prior physical hypotheses and theories, thereby guiding and constraining the discovery of PDE expressions. We have formulated this approach as a nonlinear mixed-integer programming (MIP) problem, addressed through an efficient search scheme developed for this purpose. Our experiments on newly designed Fluid Mechanics and Laser Fusion datasets demonstrate the interpretability and feasibility of this method.

69. Adam Exploits ℓ_∞-geometry of Loss Landscape via Coordinate-wise Adaptivity

链接：https://iclr.cc/virtual/2025/poster/29751 abstract： Adam outperforms SGD when training language models. Yet this advantage is not well-understood theoretically -- previous convergence analysis for Adam and SGD mainly focuses on the number of steps T and is already minimax-optimal in non-convex cases, which are both $\widetilde{O}(T^{-1/4})$. In this work, we argue that the exploitation of nice ℓ_∞-geometry is the key advantage of Adam over SGD. More specifically, we give a new convergence analysis for Adam under novel assumptions that loss is smooth under ℓ_∞-geometry rather than the more common ℓ_2-geometry, which yields a much better empirical smoothness constant for GPT-2 and ResNet models. Our experiments confirm that Adam performs much worse when the favorable ℓ_∞-geometry is changed while SGD provably remains unaffected. We also extend the convergence analysis to blockwise Adam under novel blockwise smoothness assumptions.

70. SINGER: Stochastic Network Graph Evolving Operator for High Dimensional PDEs

链接：https://iclr.cc/virtual/2025/poster/27840 abstract： We present a novel framework, Stochastlc Network Graph Evolving operatoR (SINGER), for learning the evolution operator of high-dimensional partial differential equations (PDEs). The framework uses a sub-network to approximate the solution at the initial time step and stochastically evolves the sub-network parameters over time by a graph neural network to approximate the solution at later time steps. The framework is designed to inherit the desirable properties of the parametric solution operator, including graph topology, semigroup, and stability, with a theoretical guarantee. Numerical experiments on 8 evolution PDEs of 5,10,15,20-dimensions show that our method outperforms existing baselines in almost all cases (31 out of 32), and that our method generalizes well to unseen initial conditions, equation dimensions, sub-network width, and time steps.

71. InstantPortrait: One-Step Portrait Editing via Diffusion Multi-Objective Distillation

链接：https://iclr.cc/virtual/2025/poster/29200 abstract： Real-time instruction-based portrait image editing is crucial in various applications, including filters, augmented reality, and video communications, etc. However, real-time portrait editing presents three significant challenges: identity preservation, fidelity to editing instructions, and fast model inference. Given that these aspects often present a trade-off, concurrently addressing them poses an even greater challenge. While diffusion-based image editing methods have shown promising capabilities in personalized image editing in recent years, they lack a dedicated focus on portrait editing and thus suffer from the aforementioned problems as well. To address the gap, this paper introduces an Instant-Portrait Network (IPNet), the first one-step diffusion-based model for portrait editing. We train the network in two stages. We first employ an annealing identity loss to train an Identity Enhancement Network (IDE-Net), to ensure robust identity preservation. We then train the IPNet using a novel diffusion Multi-Objective Distillation approach that integrates adversarial loss, identity distillation loss, and a novel Facial-Style Enhancing loss. The Diffusion Multi-Objective Distillation approach efficiently reduces inference steps, ensures identity consistency, and enhances the precision of instruction-based editing. Extensive comparison with prior models demonstrates IPNet as a superior model in terms of identity preservation, text fidelity, and inference speed.

72. SurFhead: Affine Rig Blending for Geometrically Accurate 2D Gaussian Surfel Head Avatars

链接：https://iclr.cc/virtual/2025/poster/31173 abstract： Recent advancements in head avatar rendering using Gaussian primitives have achieved significantly high-fidelity results. Although precise head geometry is crucial for applications like mesh reconstruction and relighting, current methods struggle to capture intricate geometric details and render unseen poses due to their reliance on similarity transformations, which cannot handle stretch and shear transforms essential for detailed deformations of geometry. To address this, we propose SurFhead, a novel method that reconstructs riggable head geometry from RGB videos using 2D Gaussian surfels, which offer well-defined geometric properties, such as precise depth from fixed ray intersections and normals derived from their surface orientation, making them advantageous over 3D counterparts. SurFhead ensures high-fidelity rendering of both normals and images, even in extreme poses, by leveraging classical mesh-based deformation transfer and affine transformation interpolation. SurFhead introduces precise geometric deformation and blends surfels through polar decomposition of transformations, including those affecting normals. Our key contribution lies in bridging classical graphics techniques, such as mesh-based deformation, with modern Gaussian primitives, achieving state-of-the-art geometry reconstruction and rendering quality. Unlike previous avatar rendering approaches, SurFhead enables efficient reconstruction driven by Gaussian primitives while preserving high-fidelity geometry.

73. MAP: Multi-Human-Value Alignment Palette

链接：https://iclr.cc/virtual/2025/poster/29882 abstract： Ensuring that generative AI systems align with human values is essential but challenging, especially when considering multiple human values and their potential trade-offs. Since human values can be personalized and dynamically change over time, the desirable levels of value alignment vary across different ethnic groups, industry sectors, and user cohorts. Within existing frameworks, it is hard to define human values and align AI systems accordingly across different directions simultaneously, such as harmlessness, helpfulness, and positiveness. To address this, we develop a novel, first-principle approach called Multi-Human-Value Alignment Palette (MAP), which navigates the alignment across multiple human values in a structured and reliable way. MAP formulates the alignment problem as an optimization task with user-defined constraints, which define human value targets. It can be efficiently solved via a primal-dual approach, which determines whether a user-defined alignment target is achievable and how to achieve it. We conduct a detailed theoretical analysis of MAP by quantifying the trade-offs between values, the sensitivity to constraints, the fundamental connection between multi-value alignment and sequential alignment, and proving that linear weighted rewards are sufficient for multi-value alignment. Extensive experiments demonstrate MAP's ability to align multiple values in a principled manner while delivering strong empirical performance across various tasks.

74. Near-Optimal Policy Identification in Robust Constrained Markov Decision Processes via Epigraph Form

链接：https://iclr.cc/virtual/2025/poster/30302 abstract： Designing a safe policy for uncertain environments is crucial in real-world control systems. However, this challenge remains inadequately addressed within the Markov decision process (MDP) framework. This paper presents the first algorithm guaranteed to identify a near-optimal policy in a robust constrained MDP (RCMDP), where an optimal policy minimizes cumulative cost while satisfying constraints in the worst-case scenario across a set of environments. We first prove that the conventional policy gradient approach to the Lagrangian max-min formulation can become trapped in suboptimal solutions. This occurs when its inner minimization encounters a sum of conflicting gradients from the objective and constraint functions. To address this, we leverage the epigraph form of the RCMDP problem, which resolves the conflict by selecting a single gradient from either the objective or the constraints. Building on the epigraph form, we propose a bisection search algorithm with a policy gradient subroutine and prove that it identifies an ε-optimal policy in an RCMDP with $\widetilde{\mathcal{O}}(\varepsilon^{-4})$ robust policy evaluations.

75. Diffusion Transformers for Tabular Data Time Series Generation

链接：https://iclr.cc/virtual/2025/poster/29095 abstract： Tabular data generation has recently attracted a growing interest due to its different application scenarios. However, generating time series of tabular data, where each element of the series depends on the others,remains a largely unexplored domain. This gap is probably due to the difficulty of jointly solving different problems, the main of which are the heterogeneity of tabular data (a problem common to non-time-dependent approaches) and the variable length of a time series.In this paper, we propose a Diffusion Transformers (DiTs) based approach for tabular data series generation. Inspired by the recent success of DiTs in image and video generation, we extend this framework to deal with heterogeneous data and variable-length sequences. Using extensive experiments on six datasets, we show that the proposed approach outperforms previous work by a large margin.

76. Optimality and Adaptivity of Deep Neural Features for Instrumental Variable Regression

链接：https://iclr.cc/virtual/2025/poster/29643 abstract： We provide a convergence analysis of \emph{deep feature instrumental variable} (DFIV) regression (Xu et al., 2021), a nonparametric approach to IV regression using data-adaptive features learned by deep neural networks in two stages. We prove that the DFIV algorithm achieves the minimax optimal learning rate when the target structural function lies in a Besov space. This is shown under standard nonparametric IV assumptions, and an additional smoothness assumption on the regularity of the conditional distribution of the covariate given the instrument, which controls the difficulty of Stage 1. We further demonstrate that DFIV, as a data-adaptive algorithm, is superior to fixed-feature (kernel or sieve) IV methods in two ways. First, when the target function possesses low spatial homogeneity (i.e., it has both smooth and spiky/discontinuous regions), DFIV still achieves the optimal rate, while fixed-feature methods are shown to be strictly suboptimal. Second, comparing with kernel-based two-stage regression estimators, DFIV is provably more data efficient in the Stage 1 samples.

77. Revisiting Large-Scale Non-convex Distributionally Robust Optimization

链接：https://iclr.cc/virtual/2025/poster/30100 abstract： Distributionally robust optimization (DRO) is a powerful technique to train robust machine learning models that perform well under distribution shifts. Compared with empirical risk minimization (ERM), DRO optimizes the expected loss under the worst-case distribution inan uncertainty set of distributions. This paper revisits the important problem of DRO with non-convex smooth loss functions. For this problem, Jin et al. (2021) showed that its dual problem is generalized (L_0, L_1)-smooth condition and gradient noise satisfies the affine variance condition, designed an algorithm of mini-batch normalized gradient descent with momentum, and proved its convergence and complexity. In this paper, we show that the dual problem and the gradient noise satisfy simpler yet more precise partially generalized smoothness

condition and partially affine variance condition by studying the optimization variable and dual variable separately, which further yields much simpler algorithm design and convergence analysis. We develop a double stochastic gradient descent with clipping (D-SGD-C) algorithm that converges to an ϵ-stationary point with $\mathcal O(\epsilon^{-4})$ gradient complexity, which matches with results in Jin et al. (2021). Our algorithm does not need to use momentum, and the proof is much simpler, thanks to the more precise characterization of partially generalized smoothness and partially affine variance noise. We further design a variance-reduced method that achieves a lower gradient complexity of $\mathcal O(\epsilon^{-3})$. Our theoretical results and insights are further verified numerically on a number of tasks, and our algorithms outperform the existing DRO method (Jin et al., 2021).

78. Mufu: Multilingual Fused Learning for Low-Resource Translation with LLM

链接：https://iclr.cc/virtual/2025/poster/31245 abstract： Multilingual large language models (LLMs) are great translators, but this is largely limited to high-resource languages. For many LLMs, translating in and out of low-resource languages remains a challenging task. To maximize data efficiency in this low-resource setting, we introduce Mufu, which includes a selection of automatically generated multilingual candidates and an instruction to correct inaccurate translations in the prompt. Mufu prompts turn a translation task into a postediting one, and seek to harness the LLM’s reasoning capability with auxiliary translation candidates, from which the model is required to assess the input quality, align the semantics cross-lingually, copy from relevant inputs and override instances that are incorrect. Our experiments on En-XX translations over the Flores-200 dataset show LLMs finetuned against Mufu-style prompts are robust to poor quality auxiliary translation candidates, achieving performance superior to NLLB 1.3B distilled model in 64% of low- and very-low-resource language pairs. We then distill these models to reduce inference cost, while maintaining on average 3.1 chrF improvement over finetune-only baseline in low-resource translations.

79. ImageFolder: Autoregressive Image Generation with Folded Tokens

链接：https://iclr.cc/virtual/2025/poster/29703 abstract： Image tokenizers are crucial for visual generative models, \eg, diffusion models (DMs) and autoregressive (AR) models, as they construct the latent representation for modeling. Increasing token length is a common approach to improve image reconstruction quality. However, tokenizers with longer token lengths are not guaranteed to achieve better generation quality. There exists a trade-off between reconstruction and generation quality regarding token length. In this paper, we investigate the impact of token length on both image reconstruction and generation and provide a flexible solution to the tradeoff. We propose \textbf{ImageFolder}, a semantic tokenizer that provides spatially aligned image tokens that can be folded during autoregressive modeling to improve both efficiency and quality. To enhance the representative capability without increasing token length, we leverage dual-branch product quantization to capture different contexts of images. Specifically, semantic regularization is introduced in one branch to encourage compacted semantic information while another branch is designed to capture pixel-level details. Extensive experiments demonstrate the superior quality of image generation and shorter token length with ImageFolder tokenizer.

80. High-Dynamic Radar Sequence Prediction for Weather Nowcasting Using Spatiotemporal Coherent Gaussian Representation

链接：https://iclr.cc/virtual/2025/poster/30500 abstract： Weather nowcasting is an essential task that involves predicting future radar echo sequences based on current observations, offering significant benefits for disaster management, transportation, and urban planning. Current prediction methods are limited by training and storage efficiency, mainly focusing on 2D spatial predictions at specific altitudes. Meanwhile, 3D volumetric predictions at each timestamp remain largely unexplored. To address such a challenge, we introduce a comprehensive framework for 3D radar sequence prediction in weather nowcasting, using the newly proposed SpatioTemporal Coherent Gaussian Splatting (STC-GS) for dynamic radar representation and GauMamba for efficient and accurate forecasting. Specifically, rather than relying on a 4D Gaussian for dynamic scene reconstruction, STC-GS optimizes 3D scenes at each frame by employing a group of Gaussians while effectively capturing their movements across consecutive frames. It ensures consistent tracking of each Gaussian over time, making it particularly effective for prediction tasks. With the temporally correlated Gaussian groups established, we utilize them to train GauMamba, which integrates a memory mechanism into the Mamba framework. This allows the model to learn the temporal evolution of Gaussian groups while efficiently handling a large volume of Gaussian tokens. As a result, it achieves both efficiency and accuracy in forecasting a wide range of dynamic meteorological radar signals. The experimental results demonstrate that our STC-GS can efficiently represent 3D radar sequences with over $16\times$ higher spatial resolution compared with the existing 3D representation methods, while GauMamba outperforms state-of-the-art methods in forecasting a broad spectrum of high-dynamic weather conditions.

81. Aligning Human Motion Generation with Human Perceptions

链接：https://iclr.cc/virtual/2025/poster/29694 abstract： Human motion generation is a critical task with a wide spectrum of applications. Achieving high realism in generated motions requires naturalness, smoothness, and plausibility. However, current evaluation metrics often rely on simple heuristics or distribution distances and do not align well with human perceptions. In this work, we propose a data-driven approach to bridge this gap by introducing a large-scale human perceptual evaluation dataset, MotionPercept, and a human motion critic model, MotionCritic, that capture human perceptual preferences. Our critic model offers a more accurate metric for assessing motion quality and could be readily integrated into the motion generation pipeline to enhance generation quality. Extensive experiments demonstrate the effectiveness of our approach in both evaluating and

improving the quality of generated human motions by aligning with human perceptions. Code and data are publicly available at https://motioncritic.github.io/.

82. SEPARATE: A Simple Low-rank Projection for Gradient Compression in Modern Large-scale Model Training Process

链接：https://iclr.cc/virtual/2025/poster/30774 abstract： Training Large Language Models (LLMs) presents a significant communication bottleneck, predominantly due to the growing scale of the gradient to communicate across multi-device clusters. However, how to mitigate communication overhead in practice remains a formidable challenge due to the weakness of the methodology of the existing compression methods, especially the neglect of the characteristics of the gradient. In this paper, we consider and demonstrate the low-rank properties of gradient and Hessian observed in LLMs training dynamic, and take advantage of such natural properties to design SEPARATE, a simple low-rank projection for gradient compression in modern large-scale model training processes. SEPARATE realizes dimensional reduction by common random Gaussian variables and an improved moving average error-feedback technique. We theoretically demonstrate that SEPARATE-based optimizers maintain the original convergence rate for SGD and Adam-Type optimizers for general non-convex objectives. Experimental results show that SEPARATE accelerates training speed by up to 2× for GPT-2-Medium pre-training, and improves performance on various benchmarks for LLAMA2-7B fine-tuning.

83. Can a MISL Fly? Analysis and Ingredients for Mutual Information Skill Learning

链接：https://iclr.cc/virtual/2025/poster/27762 abstract： Self-supervised learning has the potential of lifting several of the key challenges in reinforcement learning today, such as exploration, representation learning, and reward design. Recent work (METRA) has effectively argued that moving away from mutual information and instead optimizing a certain Wasserstein distance is important for good performance. In this paper, we argue that the benefits seen in that paper can largely be explained within the existing framework of mutual information skill learning (MISL).Our analysis suggests a new MISL method (contrastive successor features) that retains the excellent performance of METRA with fewer moving parts, and highlights connections between skill learning, contrastive representation learning, and successor features. Finally, through careful ablation studies, we provide further insight into some of the key ingredients for both our method and METRA.

84. Affine Steerable Equivariant Layer for Canonicalization of Neural Networks

链接：https://iclr.cc/virtual/2025/poster/30933 abstract： In the field of equivariant networks, achieving affine equivariance, particularly for general group representations, has long been a challenge.In this paper, we propose the steerable EquivarLayer, a generalization of InvarLayer (Li et al., 2024), by building on the concept of equivariants beyond invariants.The steerable EquivarLayer supports affine equivariance with arbitrary input and output representations, marking the first model to incorporate steerability into networks for the affine group.To integrate it with canonicalization, a promising approach for making pre-trained models equivariant, we introduce a novel Det-Pooling module, expanding the applicability of EquivarLayer and the range of groups suitable for canonicalization.We conduct experiments on image classification tasks involving group transformations to validate the steerable EquivarLayer in the role of a canonicalization function, demonstrating its effectiveness over data augmentation.

85. Three Mechanisms of Feature Learning in a Linear Network

链接：https://iclr.cc/virtual/2025/poster/29349 abstract： Understanding the dynamics of neural networks in different width regimes is crucial for improving their training and performance. We present an exact solution for the learning dynamics of a one-hidden-layer linear network, with one-dimensional data, across any finite width, uniquely exhibiting both kernel and feature learning phases. This study marks a technical advancement by enabling the analysis of the training trajectory from any initialization and a detailed phase diagram under varying common hyperparameters such as width, layer-wise learning rates, and scales of output and initialization. We identify three novel prototype mechanisms specific to the feature learning regime: (1) learning by alignment, (2) learning by disalignment, and (3) learning by rescaling, which contrast starkly with the dynamics observed in the kernel regime. Our theoretical findings are substantiated with empirical evidence showing that these mechanisms also manifest in deep nonlinear networks handling real-world tasks, enhancing our understanding of neural network training dynamics and guiding the design of more effective learning strategies.

86. Number Cookbook: Number Understanding of Language Models and How to Improve It

链接：https://iclr.cc/virtual/2025/poster/30568 abstract： Large language models (LLMs) can solve an increasing number of complex reasoning tasks while making surprising mistakes in basic numerical understanding and processing (such as $9.11 > 9.9$). The latter ability is essential for tackling complex arithmetic and mathematical problems and serves as a foundation for most reasoning tasks, but previous work paid little attention to it or only discussed several restricted tasks (like integer addition).

In this paper, we comprehensively investigate the numerical understanding and processing ability (NUPA) of LLMs. Firstly, we introduce a benchmark covering four common numerical representations and 17 distinct numerical tasks in four major categories, resulting in 41 meaningful combinations in total. These tasks are derived from primary and secondary education curricula, encompassing nearly all everyday numerical understanding and processing scenarios, and the rules of these tasks are very simple and clear.Through the benchmark, we find that current LLMs fail frequently in many of the tasks. To study the problem, we train small models with existing and potential techniques for enhancing NUPA (such as tokenizers, PEs, and number formats), comprehensively evaluating their effectiveness using our testbed. We also finetune practical-scale LLMs on our proposed NUPA tasks and find that 1) naive finetuning can improve NUPA a lot on many but not all tasks, and 2) surprisingly, techniques designed to enhance NUPA prove ineffective for finetuning pretrained models. We further explore the impact of chain-of-thought techniques on NUPA. Our work provides a more detailed and comprehensive understanding of NUPA in LLMs.

87. MAVIS: Mathematical Visual Instruction Tuning with an Automatic Data Engine

链接：https://iclr.cc/virtual/2025/poster/29918 abstract： Multi-modal Large Language Models (MLLMs) have recently showcased superior proficiency in general visual scenarios. However, we identify their mathematical capabilities remain under-explored with three areas to be improved: visual encoding of math diagrams, diagram-language alignment, and chain-of-thought (CoT) reasoning. This draws forth an urgent demand for an effective training paradigm and a large-scale, comprehensive dataset with detailed CoT rationales, which is challenging to collect and costly to annotate manually. To tackle this issue, we propose MAVIS, a MAthematical VISual instruction tuning pipeline for MLLMs, featuring an automatic data engine to efficiently create mathematical visual datasets.We design the data generation process to be entirely independent of human intervention or GPT API usage, while ensuring the diagram-caption correspondence, question-answer correctness, and CoT reasoning quality. With this approach, we curate two datasets, MAVIS-Caption (558K diagram-caption pairs) and MAVIS-Instruct (834K visual math problems with CoT rationales), and propose four progressive stages for training MLLMs from scratch.First, we utilize MAVIS-Caption to fine-tune a math-specific vision encoder (CLIP-Math) through contrastive learning, tailored for improved diagram visual encoding. Second, we also leverage MAVIS-Caption to align the CLIP-Math with a large language model (LLM) by a projection layer, enhancing vision-language alignment in mathematical domains. Third, we adopt MAVIS-Instruct to perform the instruction tuning for robust problem-solving skills, and term the resulting model as MAVIS-7B. Fourth, we apply Direct Preference Optimization (DPO) to enhance the CoT capabilities of our model, further refining its step-wise reasoning performance.On various mathematical benchmarks, our MAVIS-7B achieves leading results among open-source MLLMs, e.g., surpassing other 7B models by +9.3% and the second-best LLaVA-NeXT (110B) by +6.9%, demonstrating the effectiveness of our method.

88. TC-MoE: Augmenting Mixture of Experts with Ternary Expert Choice

链接：https://iclr.cc/virtual/2025/poster/28959 abstract： The Mixture of Experts (MoE) architecture has emerged as a promising solution to reduce computational overhead by selectively activating subsets of model parameters.The effectiveness of MoE models depends primarily on their routing mechanisms, with the widely adopted Top-K routing scheme used for activating experts.However, the Top-K scheme has notable limitations,including unnecessary activations and underutilization of experts.In this work, rather than modifying the routing mechanism as done in previous studies,we propose the Ternary Choice MoE (TC-MoE),a novel approach that expands the expert space by applying the ternary set {-1, 0, 1} to each expert.This expansion allows more efficient and effective expert activations without incurring significant computational costs.Additionally, given the unique characteristics of the expanded expert space,we introduce a new load balance loss and reward loss to ensure workload balance and achieve a flexible trade-off between effectiveness and efficiency.Extensive experiments demonstrate that TC-MoE achieves an average improvement of over 1.1% compared with traditional approaches,while reducing the average number of activated experts by up to 9%.These results confirm that TC-MoE effectively addresses the inefficiencies of conventional routing schemes,offering a more efficient and scalable solution for MoE-based large language models.Code and models are available at https://github.com/stiger1000/TC-MoE.

89. Provably Accurate Shapley Value Estimation via Leverage Score Sampling

链接：https://iclr.cc/virtual/2025/poster/27831 abstract： Originally introduced in game theory, Shapley values have emerged as a central tool in explainable machine learning, where they are used to attribute model predictions to specific input features. However, computing Shapley values exactly is expensive: for a model with n features, $O(2^n)$ model evaluations are necessary. To address this issue, approximation algorithms are widely used. One of the most popular is the Kernel SHAP algorithm, which is model agnostic and remarkably effective in practice. However, to the best of our knowledge, Kernel SHAP has no strong non-asymptotic complexity guarantees. We address this issue by introducing *Leverage SHAP*, a light-weight modification of Kernel SHAP that provides provably accurate Shapley value estimates with just $O(n\log n)$ model evaluations. Our approach takes advantage of a connection between Shapley value estimation and agnostic active learning by employing *leverage score sampling*, a powerful regression tool. Beyond theoretical guarantees, we show that Leverage SHAP consistently outperforms even the highly optimized implementation of Kernel SHAP available in the ubiquitous SHAP library [Lundberg \& Lee, 2017].

90. Remove Symmetries to Control Model Expressivity and Improve

Optimization

链接：https://iclr.cc/virtual/2025/poster/30250 abstract： When symmetry is present in the loss function, the model is likely to be trapped in a low-capacity state that is sometimes known as a ``collapse." Being trapped in these low-capacity states can be a major obstacle to training across many scenarios where deep learning technology is applied. We first prove two concrete mechanisms through which symmetries lead to reduced capacities and ignored features during training and inference. We then propose a simple and theoretically justified algorithm, \textit{syre}, to remove almost all symmetry-induced low-capacity states in neural networks. When this type of entrapment is especially a concern, removing symmetries with the proposed method is shown to correlate well with improved optimization or performance. A remarkable merit of the proposed method is that it is model-agnostic and does not require any knowledge of the symmetry.

91. Attention with Markov: A Curious Case of Single-layer Transformers

链接：https://iclr.cc/virtual/2025/poster/29578 abstract： Attention-based transformers have achieved tremendous success across a variety of disciplines including natural languages. To deepen our understanding of their sequential modeling capabilities, there is a growing interest in using Markov input processes to study them. A key finding is that when trained on first-order Markov chains, transformers with two or more layers consistently develop an induction head mechanism to estimate the in-context bigram conditional distribution. In contrast, single-layer transformers, unable to form an induction head, directly learn the Markov kernel but often face a surprising challenge: they become trapped in local minima representing the unigram distribution, whereas deeper models reliably converge to the ground-truth bigram. While single-layer transformers can theoretically model first-order Markov chains, their empirical failure to learn this simple kernel in practice remains a curious phenomenon. To explain this contrasting behavior of single-layer models, in this paper we introduce a new framework for a principled analysis of transformers via Markov chains. Leveraging our framework, we theoretically characterize the loss landscape of single-layer transformers and show the existence of global minima (bigram) and bad local minima (unigram) contingent on data properties and model architecture. We precisely delineate the regimes under which these local optima occur. Backed by experiments, we demonstrate that our theoretical findings are in congruence with the empirical results. Finally, we outline several open problems in this arena.

92. The AdEMAMix Optimizer: Better, Faster, Older

链接：https://iclr.cc/virtual/2025/poster/28625 abstract： Momentum based optimizers are central to a wide range of machine learning applications. These typically rely on an Exponential Moving Average (EMA) of gradients, which decays exponentially the present contribution of older gradients. This accounts for gradients being local linear approximations which lose their relevance as the iterate moves along the loss landscape. This work questions the use of a single EMA to accumulate past gradients and empirically demonstrates how this choice can be sub-optimal: a single EMA cannot simultaneously give a high weight to the immediate past, and a non-negligible weight to older gradients. Building on this observation, we propose AdEMAMix, a simple modification of the Adam optimizer with a mixture of two EMAs to better take advantage of past gradients. Our experiments on language modeling and image classification show---quite surprisingly---that gradients can stay relevant for tens of thousands of steps. They help to converge faster, and often to lower minima: e.g., a 1.3B parameter AdEMAMix LLM trained on 101B tokens performs comparably to an AdamW model trained on 197B tokens ($+95\%$). Moreover, our method significantly slows-down model forgetting during training. Our work motivates further exploration of different types of functions to leverage past gradients, beyond EMAs.

93. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code

链接：https://iclr.cc/virtual/2025/poster/29033 abstract： Large Language Models (LLMs) applied to code-related applications have emerged as a prominent field, attracting significant interest from academia and industry. However, as new and improved LLMs are developed, existing evaluation benchmarks (e.g., HumanEvla, MBPP) are no longer sufficient for assessing their capabilities suffering from data contamination, overfitting, saturation, and focus on merely code generation. In this work, we propose LiveCodeBench, a comprehensive and contamination-free evaluation of LLMs for code, which collects new problems over time from contests across three competition platforms, Leetcode, Atcoder, and Codeforces. Notably, our benchmark also focuses on a broader range of code-related capabilities, such as self-repair, code execution, and test output prediction, beyond just code generation. Currently, LiveCodeBench hosts over six hundred coding problems that were published between May 2023 and Aug 2024. We evaluate over 50 LLMs on LiveCodeBench (LCB for brevity) presenting the largest evaluation study of code LLMs on competition problems. Based on the study, we present novel empirical findings on contamination, overfitting, and holistic evaluations. We demonstrate that time-segmented evaluations serve as a robust approach to evade contamination; they are successful at detecting contamination across a wide range of open and closed models including GPT-4O, Claude, Deepseek, and Codestral. Next, we highlight overfitting and saturation of traditional coding benchmarks like HumanEvla and demonstrate LCB allows more reliable evaluations. Finally, our holistic evaluation scenarios allow for measuring the different capabilities of programming agents in isolation.

94. Pyramidal Flow Matching for Efficient Video Generative Modeling

链接：https://iclr.cc/virtual/2025/poster/30909 abstract： Video generation requires modeling a vast spatiotemporal space,

which demands significant computational resources and data usage. To reduce the complexity, the prevailing approaches employ a cascaded architecture to avoid direct training with full resolution latent. Despite reducing computational demands, the separate optimization of each sub-stage hinders knowledge sharing and sacrifices flexibility. This work introduces a unified pyramidal flow matching algorithm. It reinterprets the original denoising trajectory as a series of pyramid stages, where only the final stage operates at the full resolution, thereby enabling more efficient video generative modeling. Through our sophisticated design, the flows of different pyramid stages can be interlinked to maintain continuity. Moreover, we craft autoregressive video generation with a temporal pyramid to compress the full-resolution history. The entire framework can be optimized in an end-to-end manner and with a single unified Diffusion Transformer (DiT). Extensive experiments demonstrate that our method supports generating high-quality 5-second (up to 10-second) videos at 768p resolution and 24 FPS within 20.7k A100 GPU training hours. All code and models are open-sourced at https://pyramid-flow.github.io.

95. DeLLMa: Decision Making Under Uncertainty with Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30624 abstract： The potential of large language models (LLMs) as decision support tools is increasingly being explored in fields such as business, engineering, and medicine, which often face challenging tasks of decision-making under uncertainty. In this paper, we show that directly prompting LLMs on these types of decision-making problems can yield poor results, especially as the problem complexity increases. To aid in these tasks, we propose DeLLMa (Decision-making Large Language Model assistant), a framework designed to enhance decision-making accuracy in uncertain environments. DeLLMa involves a multi-step reasoning procedure that integrates recent best practices in scaling inference-time reasoning, drawing upon principles from decision theory and utility theory, to provide an accurate and human-auditable decision-making process. We validate our procedure on multiple realistic decision-making environments, demonstrating that DeLLMa can consistently enhance the decision-making performance of leading language models, and achieve up to a 40% increase in accuracy over competing methods. Additionally, we show how performance improves when scaling compute at test time, and carry out human evaluations to benchmark components of DeLLMa.

96. JPEG Inspired Deep Learning

链接：https://iclr.cc/virtual/2025/poster/28038 abstract： Although it is traditionally believed that lossy image compression, such as JPEG compression, has a negative impact on the performance of deep neural networks (DNNs), it is shown by recent works that well-crafted JPEG compression can actually improve the performance of deep learning (DL). Inspired by this, we propose JPEG-DL, a novel DL framework that prepends any underlying DNN architecture with a trainable JPEG compression layer. To make the quantization operation in JPEG compression trainable, a new differentiable soft quantizer is employed at the JPEG layer, and then the quantization operation and underlying DNN are jointly trained. Extensive experiments show that in comparison with the standard DL, JPEG-DL delivers significant accuracy improvements across various datasets and model architectures while enhancing robustness against adversarial attacks. Particularly, on some fine-grained image classification datasets, JPEG-DL can increase prediction accuracy by as much as 20.9%. Our code is available on https://github.com/AhmedHussKhalifa/JPEG-Inspired-DL.git.

97. SpinQuant: LLM Quantization with Learned Rotations

链接：https://iclr.cc/virtual/2025/poster/28338 abstract： Post-training quantization (PTQ) techniques applied to weights, activations, and the KV cache greatly reduce memory usage, latency, and power consumption of Large Language Models (LLMs), but may lead to large quantization errors when outliers are present. Rotating activation or weight matrices helps remove outliers and benefits quantization. In this work, we identify a collection of applicable rotation parameterizations that lead to identical outputs in full-precision Transformer architectures while enhancing quantization accuracy. In addition, we find that some random rotations lead to much better quantization than others, with an up to 13 points difference in downstream zero-shot reasoning performance. As a result, we propose SpinQuant, a novel approach that incorporates learned rotation matrices for optimal quantized network accuracy. With 4-bit quantization of weight, activation, and KV-cache, SpinQuant narrows the accuracy gap on zero-shot reasoning tasks with full precision to merely 2.9 points on the LLaMA-2 7B model, surpassing LLM-QAT by 19.1 points and SmoothQuant by 25.0 points. Furthermore, SpinQuant also outperforms concurrent work QuaRot, which applies random rotations to remove outliers. In particular, for LLaMA-3 8B models that are hard to quantize, SpinQuant reduces the gap to full precision by up to 45.1% relative to QuaRot. Code is available at https://github.com/facebookresearch/SpinQuant.

98. R-Sparse: Rank-Aware Activation Sparsity for Efficient LLM Inference

链接：https://iclr.cc/virtual/2025/poster/30686 abstract： Large Language Models (LLMs), while demonstrating remarkable capabilities across various applications, present significant challenges during inference due to their substantial model size, especially when deployed on edge devices. Activation sparsity offers a promising solution to reduce computation and memory movement, enabling more efficient inference, particularly for small-batch on-device applications. However, current approaches face limitations with non-ReLU activation function, which are foundational to most advanced LLMs, or require heavy continual training. Additionally, the difficulty in predicting active channels and limited achievable sparsity ratios constrain the effectiveness of activation sparsity-based methods. In this paper, we introduce R-Sparse, a training-free activation sparsity approach capable of achieving high sparsity levels in advanced LLMs. We conducted two preliminary investigations into how different components contribute to the output within a single linear layer and found two key observations: (i) the non-sparse components of the input function can be regarded as a few bias terms, and (ii) The full computation can be effectively approximated by an appropriate

combination of input channels and weight singular values. Building on this, we replace the linear layers in LLMs with a rank-aware sparse inference method that leverages the sparsity of input channels and singular value components, eliminating the need for active channel prediction like the output sparsity based approaches. Experiments on Llama-2/3 and Mistral models across ten diverse tasks demonstrate that R-Sparse achieves comparable performance at 50\% model-level sparsity, resulting in a significant 43\% end-to-end efficient improvements with customized kernels.

99. ParamΔ for Direct Mixing: Post-Train Large Language Model At Zero Cost

链接：https://iclr.cc/virtual/2025/poster/27886 abstract： The post-training phase of large language models is essential for enhancing capabilities such as instruction-following, reasoning, and alignment with human preferences. However, it demands extensive high-quality data and poses risks like overfitting, alongside significant computational costs due to repeated post-training and evaluation after each base model update. This paper introduces ParamΔ, a novel method that streamlines post-training by transferring knowledge from an existing post-trained model to a newly updated base model with \textbf{zero} additional training. By computing the difference between post-trained model weights (Θ_post) and base model weights (Θ_base), and adding this to the updated base model (Θ_base'), we define ParamΔ Model as: $\Theta_{\text{Param}\Delta} = \Theta_\text{post} - \Theta_\text{base} + \Theta_\text{base}'$. This approach surprisingly equips the new base model with post-trained capabilities, achieving performance comparable to direct post-training. We did analysis on LLama3, Llama3.1, Qwen, and DeepSeek-distilled models. Results indicate ParamΔ Model effectively replicates traditional post-training. For example, the ParamΔ Model obtained from 70B Llama3-inst, Llama3-base, Llama3.1-base models attains approximately 95\% of Llama3.1-inst model's performance on average. ParamΔ brings a new perspective on how to fully leverage models in the open-weight community, where checkpoints for base and instruct models are readily available and frequently updated, by providing a cost-free framework to accelerate the iterative cycle of model development.

100. ChatQA 2: Bridging the Gap to Proprietary LLMs in Long Context and RAG Capabilities

链接：https://iclr.cc/virtual/2025/poster/29053 abstract： In this work, we introduce ChatQA 2, an Llama 3.0-based model with a 128Kcontext window, designed to bridge the gap between open-source LLMs andleading proprietary models (e.g., GPT-4-Turbo-2024-04-09) in long context un-derstanding and retrieval-augmented generation (RAG) capabilities. These twocapabilities are complementary to each other and essential for LLMs to processlarge volumes of information that cannot fit into a single prompt. We presenta detailed continued training recipe to extend the context window of Llama3-70B-base from 8K to 128K tokens, along with a three-stage instruction tun-ing process to enhance the model’s instruction-following, RAG performance,and long-context understanding capabilities. Our results demonstrate that theLlama3-ChatQA-2-70B model outperforms most existing state-of-the-art models,including GPT-4-Turbo-2024-04-09, Qwen2-72B-Instruct, and Llama3.1-70B-Instruct, on ultra-long tasks beyond 100K tokens, as well as on the RAG benchmarkusing only a 4K context window, showing the strong long context capability acrossvarying sequence lengths. We further provide extensive comparisons betweendirect long-context and RAG solutions using the same state-of-the-art long-contextLLMs. Interestingly, we find that the performance of strong long-context LLMsusing RAG improves when retrieving a larger number of chunks. With a large setof top-k chunks, RAG consistently outperforms direct long-context solution usingthe same state-of-the-art long-context models (e.g., Llama3-ChatQA-2-70B andQwen2-72B-Instruct) on both 32K and 128K benchmarks. We open-source themodel weights, training data, and the evaluation setup for the for the community:https://chatqa2-project.github.io/

101. Image-level Memorization Detection via Inversion-based Inference Perturbation

链接：https://iclr.cc/virtual/2025/poster/27880 abstract： Recent studies have discovered that widely used text-to-image diffusion models can replicate training samples during image generation, a phenomenon known as memorization. Existing detection methods primarily focus on identifying memorized prompts. However, in real-world scenarios, image owners may need to verify whether their proprietary or personal images have been memorized by the model, even in the absence of paired prompts or related metadata. We refer to this challenge as image-level memorization detection, where current methods relying on original prompts fall short. In this work, we uncover two characteristics of memorized images after perturbing the inference procedure: lower similarity of the original images and larger magnitudes of TCNP.Building on these insights, we propose Inversion-based Inference Perturbation (IIP), a new framework for image-level memorization detection. Our approach uses unconditional DDIM inversion to derive latent codes that contain core semantic information of original images and optimizes random prompt embeddings to introduce effective perturbation. Memorized images exhibit distinct characteristics within the proposed pipeline, providing a robust basis for detection. To support this task, we construct a comprehensive setup for the image-level memorization detection, carefully curating datasets to simulate realistic memorization scenarios. Using this setup, we evaluate our IIP framework across three different memorization settings, demonstrating its state-of-the-art performance in identifying memorized images in various settings, even in the presence of data augmentation attacks.

102. gRNAde: Geometric Deep Learning for 3D RNA inverse design

链接：https://iclr.cc/virtual/2025/poster/28493 abstract： Computational RNA design tasks are often posed as inverse problems, where sequences are designed based on adopting a single desired secondary structure without considering 3D conformational diversity. We introduce gRNAde, a geometric RNA design pipeline operating on 3D RNA backbones to design sequences that explicitly account for structure and dynamics. gRNAde uses a multi-state Graph Neural Network and autoregressive decoding to generates candidate RNA sequences conditioned on one or more 3D backbone structures where the identities of the bases are unknown. On a single-state fixed backbone re-design benchmark of 14 RNA structures from the PDB identified by Das et al. (2010), gRNAde obtains higher native sequence recovery rates (56% on average) compared to Rosetta (45% on average), taking under a second to produce designs compared to the reported hours for Rosetta. We further demonstrate the utility of gRNAde on a new benchmark of multi-state design for structurally flexible RNAs, as well as zero-shot ranking of mutational fitness landscapes in a retrospective analysis of a recent ribozyme. Experimental wet lab validation on 10 different structured RNA backbones finds that gRNAde has a success rate of 50% at designing pseudoknotted RNA structures, a significant advance over 35% for Rosetta. Open source code and tutorials are available at: github.com/chaitjo/geometric-rna-design

103. A Policy-Gradient Approach to Solving Imperfect-Information Games with Best-Iterate Convergence

链接：https://iclr.cc/virtual/2025/poster/29215 abstract： Policy gradient methods have become a staple of any single-agent reinforcement learning toolbox, due to their combination of desirable properties: iterate convergence, efficient use of stochastic trajectory feedback, and theoretically-sound avoidance of importance sampling corrections. In multi-agent imperfect-information settings (extensive-form games), however, it is still unknown whether the same desiderata can be guaranteed while retaining theoretical guarantees. Instead, sound methods for extensive-form games rely on approximating \emph{counterfactual} values (as opposed to Q values), which are incompatible with policy gradient methodologies. In this paper, we investigate whether policy gradient can be safely used in two-player zero-sum imperfect-information extensive-form games (EFGs). We establish positive results, showing for the first time that a policy gradient method leads to provable best-iterate convergence to a regularized Nash equilibrium in self-play.

104. Boltzmann priors for Implicit Transfer Operators

链接：https://iclr.cc/virtual/2025/poster/28298 abstract： Accurate prediction of thermodynamic properties is essential in drug discovery and materials science. Molecular dynamics (MD) simulations provide a principled approach to this task, yet they typically rely on prohibitively long sequential simulations. Implicit Transfer Operator (ITO) Learning offers a promising approach to address this limitation by enabling stable simulation with time steps orders of magnitude larger than MD. However, to train ITOs, we need extensive, unbiased MD data, limiting the scope of this framework. Here, we introduce Boltzmann Priors for ITO (BoPITO) to enhance ITO learning in two ways. First, BoPITO enables more efficient data generation, and second, it embeds inductive biases for long-term dynamical behavior, simultaneously improving sample efficiency by one order of magnitude and guaranteeing asymptotically unbiased equilibrium statistics. Furthermore, we showcase the use of BoPITO in a new tunable sampling protocol interpolating between ITOs trained on off-equilibrium simulations and an equilibrium model by incorporating unbiased correlation functions. Code is available at https://github.com/olsson-group/bopito.

105. Conformal Generative Modeling with Improved Sample Efficiency through Sequential Greedy Filtering

链接：https://iclr.cc/virtual/2025/poster/31189 abstract： Generative models lack rigorous statistical guarantees with respect to their predictions. In this work, we propose Sequential Conformal Prediction for Generative Models (SCOPE-Gen), a sequential conformal prediction method producing prediction sets that satisfy a rigorous statistical guarantee called conformal admissibility control. This guarantee means that the prediction sets contain at least one admissible (or valid) example, with high probability. To this end, our method first samples an initial set of i.i.d. examples from a black box generative model. Then, this set is iteratively pruned via so-called greedy filters. As a consequence of the iterative generation procedure, admissibility of the final prediction set factorizes as a Markov chain, where each factor can be controlled separately, using conformal prediction. In comparison to prior work, our method demonstrates a large reduction in the number of admissibility evaluations during calibration. This is crucial e.g. in safety-critical applications, where these evaluations must be conducted manually by domain experts and are therefore costly and time consuming. We highlight the advantages of our method in terms of admissibility evaluations and cardinality of the prediction set through experiments in natural language generation and molecular graph extension tasks.

106. How Does Vision-Language Adaptation Impact the Safety of Vision Language Models?

链接：https://iclr.cc/virtual/2025/poster/28921 abstract： Vision-Language adaptation (VL adaptation) transforms Large Language Models (LLMs) into Large Vision-Language Models (LVLMs) for multimodal tasks, but this process often compromises the inherent safety capabilities embedded in the original LLMs. Despite potential harmfulness due to weakened safety measures, in-depth analysis on the effects of VL adaptation on safety remains under-explored. This study examines how VL adaptation influences safety and evaluates the impact of safety fine-tuning methods. Our analysis reveals that safety

degradation occurs during VL adaptation, even when the training data is safe. While safety tuning techniques like supervised fine-tuning with safety datasets or reinforcement learning from human feedback mitigate some risks, they still lead to safety degradation and a reduction in helpfulness due to over-rejection issues. Further analysis of internal model weights suggests that VL adaptation may impact certain safety-related layers, potentially lowering overall safety levels. Additionally, our findings demonstrate that the objectives of VL adaptation and safety tuning are divergent, which often results in their simultaneous application being suboptimal. To address this, we suggest the weight merging approach as an optimal solution effectively reducing safety degradation while maintaining helpfulness. These insights help guide the development of more reliable and secure LVLMs for real-world applications.

107. Neuron-based Multifractal Analysis of Neuron Interaction Dynamics in Large Models

链接：https://iclr.cc/virtual/2025/poster/28389 abstract： In recent years, there has been increasing attention on the capabilities of large-scale models, particularly in handling complex tasks that small-scale models are unable to perform. Notably, large language models (LLMs) have demonstrated ``intelligent'' abilities such as complex reasoning and abstract language comprehension, reflecting cognitive-like behaviors. However, current research on emergent abilities in large models predominantly focuses on the relationship between model performance and size, leaving a significant gap in the systematic quantitative analysis of the internal structures and mechanisms driving these emergent abilities. Drawing inspiration from neuroscience research on brain network structure and self-organization, we propose (i) a general network representation of large models, (ii) a new analytical framework — Neuron-based Multifractal Analysis (NeuroMFA) - for structural analysis, and (iii) a novel structure-based metric as a proxy for emergent abilities of large models. By linking structural features to the capabilities of large models, NeuroMFA provides a quantitative framework for analyzing emergent phenomena in large models. Our experiments show that the proposed method yields a comprehensive measure of the network's evolving heterogeneity and organization, offering theoretical foundations and a new perspective for investigating emergence in large models.

108. Pursuing Feature Separation based on Neural Collapse for Out-of-Distribution Detection

链接：https://iclr.cc/virtual/2025/poster/28462 abstract： In the open world, detecting out-of-distribution (OOD) data, whose labels are disjoint with those of in-distribution (ID) samples, is important for reliable deep neural networks (DNNs). To achieve better detection performance, one type of approach proposes to fine-tune the model with auxiliary OOD datasets to amplify the difference between ID and OOD data through a separation loss defined on model outputs. However, none of these studies consider enlarging the feature disparity, which should be more effective compared to outputs. The main difficulty lies in the diversity of OOD samples, which makes it hard to describe their feature distribution, let alone design losses to separate them from ID features. In this paper, we neatly fence off the problem based on an aggregation property of ID features named Neural Collapse (NC). NC means that the penultimate features of ID samples within a class are nearly identical to the last layer weight of the corresponding class. Based on this property, we propose a simple but effective loss called Separation Loss, which binds the features of OOD data in a subspace orthogonal to the principal subspace of ID features formed by NC. In this way, the features of ID and OOD samples are separated by different dimensions. By optimizing the feature separation loss rather than purely enlarging output differences, our detection achieves SOTA performance on CIFAR10, CIFAR100 and ImageNet benchmarks without any additional data augmentation or sampling, demonstrating the importance of feature separation in OOD detection. Code is availableat https://github.com/Wuyingwen/Pursuing-Feature-Separation-for-OOD-Detection.

109. CLoSD: Closing the Loop between Simulation and Diffusion for multi-task character control

链接：https://iclr.cc/virtual/2025/poster/28291 abstract： Motion diffusion models and Reinforcement Learning (RL) based control for physics-based simulations have complementary strengths for human motion generation. The former is capable of generating a wide variety of motions, adhering to intuitive control such as text, while the latter offers physically plausible motion and direct interaction with the environment. In this work, we present a method that combines their respective strengths. CLoSD is a text-driven RL physics-based controller, guided by diffusion generation for various tasks. Our key insight is that motion diffusion can serve as an on-the-fly universal planner for a robust RL controller. To this end, CLoSD maintains a closed-loop interaction between two modules — a Diffusion Planner (DiP), and a tracking controller. DiP is a fast-responding autoregressive diffusion model, controlled by textual prompts and target locations, and the controller is a simple and robust motion imitator that continuously receives motion plans from DiP and provides feedback from the environment. CLoSD is capable of seamlessly performing a sequence of different tasks, including navigation to a goal location, striking an object with a hand or foot as specified in a text prompt, sitting down, and getting up.

110. Hidden in the Noise: Two-Stage Robust Watermarking for Images

链接：https://iclr.cc/virtual/2025/poster/28502 abstract： As the quality of image generators continues to improve, deepfakes become a topic of considerable societal debate. Image watermarking allows responsible model owners to detect and label their AI-generated content, which can mitigate the harm. Yet, current state-of-the-art methods in image watermarking remain vulnerable to forgery and removal attacks. This vulnerability occurs in part because watermarks distort the distribution of

generated images, unintentionally revealing information about the watermarking techniques.In this work, we first demonstrate a distortion-free watermarking method for images, based on a diffusion model's initial noise.However, detecting the watermark requires comparing the initial noise reconstructed for an image to all previously used initial noises. To mitigate these issues, we propose a two-stage watermarking framework for efficient detection. During generation, we augment the initial noise with generated Fourier patterns to embed information about the group of initial noises we used. For detection, we (i) retrieve the relevant group of noises, and (ii) search within the given group for an initial noise that might match our image. This watermarking approach achieves state-of-the-art robustness to forgery and removal against a large battery of attacks. The project code is available at https://github.com/Kasraarabi/Hidden-in-the-Noise.

111. Biologically Constrained Barrel Cortex Model Integrates Whisker Inputs and Replicates Key Brain Network Dynamics

链接：https://iclr.cc/virtual/2025/poster/29441 abstract： The brain's ability to transform sensory inputs into motor functions is central to neuroscience and crucial for the development of embodied intelligence. Sensory-motor integration involves complex neural circuits, diverse neuronal types, and intricate intercellular connections. Bridging the gap between biological realism and behavioral functionality presents a formidable challenge. In this study, we focus on the columnar structure of the superficial layers of mouse barrel cortex as a model system. We constructed a model comprising 4,218 neurons across 13 neuronal subtypes, with neural distribution and connection strengths constrained by anatomical experimental findings. A key innovation of our work is the development of an effective construction and training pipeline tailored for this biologically constrained model. Additionally, we converted an existing simulated whisker sweep dataset into a spiking-based format, enabling our network to be trained and tested on neural signals that more closely mimic those observed in biological systems. The results of object discrimination utilizing whisker signals demonstrate that our barrel cortex model, grounded in biological constraints, achieves a classification accuracy exceeds classical convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory networks (LSTMs), by an average of 8.6%, and is on par with recent spiking neural networks (SNNs) in performance. Interestingly, a whisker deprivation experiment, designed in accordance with neuroscience practices, further validates the perceptual capabilities of our model in behavioral tasks.Critically, it offers significant biological interpretability: post-training analysis reveals that neurons within our model exhibit firing characteristics and distribution patterns similar to those observed in the actual neuronal systems of the barrel cortex. This study advances our understanding of neural processing in the barrel cortex and exemplifies how integrating detailed biological structures into neural network models can enhance both scientific inquiry and artificial intelligence applications. The code is available at https://github.com/fun0515/RSNN_bfd.

112. LoRA-Pro: Are Low-Rank Adapters Properly Optimized?

链接：https://iclr.cc/virtual/2025/poster/28819 abstract： Low-rank adaptation, also known as LoRA, has emerged as a prominent method for parameter-efficient fine-tuning of foundation models.Despite its computational efficiency, LoRA still yields inferior performance compared to full fine-tuning.In this paper, we first uncover a fundamental connection between the optimization processes of LoRA and full fine-tuning: using LoRA for optimization is mathematically equivalent to full fine-tuning using a low-rank gradient for parameter updates.And this low-rank gradient can be expressed in terms of the gradients of the two low-rank matrices in LoRA.Leveraging this insight, we introduce LoRA-Pro, a method that enhances LoRA's performance by strategically adjusting the gradients of these low-rank matrices.This adjustment allows the low-rank gradient to more accurately approximate the full fine-tuning gradient, thereby narrowing the performance gap between LoRA and full fine-tuning.Furthermore, we theoretically derive the optimal solutions for adjusting the gradients of the low-rank matrices, applying them during fine-tuning in LoRA-Pro.We conduct extensive experiments across natural language understanding, dialogue generation, mathematical reasoning, code generation, and image classification tasks, demonstrating that LoRA-Pro substantially improves LoRA's performance, effectively narrowing the gap with full fine-tuning.Our code is publicly available at https://github.com/mrflogs/LoRA-Pro.

113. Proximal Mapping Loss: Understanding Loss Functions in Crowd Counting & Localization

链接：https://iclr.cc/virtual/2025/poster/30798 abstract： Crowd counting and localization involve extracting the number and distribution of crowds from images or videos using computer vision techniques. Most counting methods are based on density regression and are based on an ``intersection'' hypothesis, i.e., one pixel is influenced by multiple points in the ground truth, which is inconsistent with reality since one pixel would not contain two objects. This paper proposes Proximal Mapping Loss (PML), a density regression method that eliminates this hypothesis. {PML} divides the predicted density map into multiple point-neighbor cases through the nearest neighbor, and then dynamically constructs a learning target for each sub-case via proximal mapping, leading to more robust and accurate training. {Furthermore}, PML is theoretically linked to various existing loss functions, such as Gaussian-blurred L2 loss, Bayesian loss, and the training schemes in P2PNet and DMC, demonstrating its versatility and adaptability. Experimentally, PML significantly improves the performance of crowd counting and localization, and illustrates the robustness against annotation noise. The code is available at https://github.com/Elin24/pml.

114. Pairwise Elimination with Instance-Dependent Guarantees for Bandits with Cost Subsidy

链接：https://iclr.cc/virtual/2025/poster/28945 abstract： Multi-armed bandits (MAB) are commonly used in sequential online decision-making when the reward of each decision is an unknown random variable. In practice, however, the typical goal of maximizing total reward may be less important than minimizing the total cost of the decisions taken, subject to a reward constraint. For example, we may seek to make decisions that have at least the reward of a reference ``default'' decision. This problem was recently introduced in the Multi-Armed Bandits with Cost Subsidy (MAB-CS) framework. MAB-CS is broadly applicable to problem domains where a primary metric (cost) is constrained by a secondary metric (reward), and there is an inability to explicitly determine the trade-off between these metrics. In our work, we first introduce the Pairwise-Elimination algorithm for a simplified variant of the cost subsidy problem with a known reference arm. We then generalize PE to PE-CS to solve the MAB-CS problem in the setting where the reference arm is the un-identified optimal arm. Next, we analyze the performance of both PE and PE-CS on the dual metrics of Cost and Quality Regret. Our instance-dependent analysis of PE and PE-CS reveals that both algorithms have an order-wise logarithmic upper bound on Cost and Quality Regret, making our policy the first with such a guarantee. Finally, experiments are conducted using the MovieLens 25M dataset for both PE and PE-CS and using a synthetic toy experiment for PE-CS revealing that our method invariably outperforms the ETC-CS baseline from the literature.

115. MMKE-Bench: A Multimodal Editing Benchmark for Diverse Visual Knowledge

链接：https://iclr.cc/virtual/2025/poster/27931 abstract： Knowledge editing techniques have emerged as essential tools for updating the factual knowledge of large language models (LLMs) and multimodal models (LMMs), allowing them to correct outdated or inaccurate information without retraining from scratch. However, existing benchmarks for multimodal knowledge editing primarily focus on entity-level knowledge represented as simple triplets, which fail to capture the complexity of real-world multimodal information. To address this issue, we introduce MMKE-Bench, a comprehensive MultiModal Knowledge Editing Benchmark, designed to evaluate the ability of LMMs to edit diverse visual knowledge in real-world scenarios. MMKE-Bench addresses these limitations by incorporating three types of editing tasks: visual entity editing, visual semantic editing, and user-specific editing. Besides, MMKE-Bench uses free-form natural language to represent and edit knowledge, offering a more flexible and effective format. The benchmark consists of 2,940 pieces of knowledge and 8,363 images across 33 broad categories, with evaluation questions automatically generated and human-verified. We assess five state-of-the-art knowledge editing methods on three prominent LMMs, revealing that no method excels across all criteria, and that visual and user-specific edits are particularly challenging. MMKE-Bench sets a new standard for evaluating the robustness of multimodal knowledge editing techniques, driving progress in this rapidly evolving field.

116. In-Context Editing: Learning Knowledge from Self-Induced Distributions

链接：https://iclr.cc/virtual/2025/poster/27872 abstract： In scenarios where language models must incorporate new information efficiently without extensive retraining, traditional fine-tuning methods are prone to overfitting, degraded generalization, and unnatural language generation. To address these limitations, we introduce Consistent In-Context Editing (ICE), a novel approach leveraging the model's in-context learning capability to optimize towards a contextual distribution rather than a one-hot target. ICE introduces a simple yet effective optimization framework for the model to internalize new knowledge by aligning its output distributions with and without additional context. This method enhances the robustness and effectiveness of gradient-based tuning methods, preventing overfitting and preserving the model's integrity. We analyze ICE across four critical aspects of knowledge editing: accuracy, locality, generalization, and linguistic quality, demonstrating its advantages. Experimental results confirm the effectiveness of ICE and demonstrate its potential for continual editing, ensuring that the integrity of the model is preserved while updating information.

117. AutoDAN-Turbo: A Lifelong Agent for Strategy Self-Exploration to Jailbreak LLMs

链接：https://iclr.cc/virtual/2025/poster/29096 abstract： Jailbreak attacks serve as essential red-teaming tools, proactively assessing whether LLMs can behave responsibly and safely in adversarial environments. Despite diverse strategies (e.g., cipher, low-resource language, persuasions, and so on) that have been proposed and shown success, these strategies are still manually designed, limiting their scope and effectiveness as a red-teaming tool. In this paper, we propose AutoDAN-Turbo, a black-box jailbreak method that can automatically discover as many jailbreak strategies as possible from scratch, without any human intervention or predefined scopes (e.g., specified candidate strategies), and use them for red-teaming. As a result, AutoDAN-Turbo can significantly outperform baseline methods, achieving a 74.3% higher average attack success rate on public benchmarks. Notably, AutoDAN-Turbo achieves an 88.5 attack success rate on GPT-4-1106-turbo. In addition, AutoDAN-Turbo is a unified framework that can incorporate existing human-designed jailbreak strategies in a plug-and-play manner. By integrating human-designed strategies, AutoDAN-Turbo can even achieve a higher attack success rate of 93.4 on GPT-4-1106-turbo.

118. Towards Understanding the Universality of Transformers for Next-Token Prediction

链接：https://iclr.cc/virtual/2025/poster/27720 abstract： Causal Transformers are trained to predict the next token for a given context. While it is widely accepted that self-attention is crucial for encoding the causal structure of sequences, the precise underlying mechanism behind this in-context autoregressive learning ability remains unclear. In this paper, we take a step towards understanding this phenomenon by studying the approximation ability of Transformers for next-token prediction. Specifically, we explore the capacity of causal Transformers to predict the next token x_{t+1} given an autoregressive sequence $(x_1, \dots, x_t)$ as a prompt, where $ x_{t+1} = f(x_t) $, and $ f $ is a context-dependent function that varies with each sequence.On the theoretical side, we focus on specific instances, namely when $ f $ is linear or when $ (x_t)$ is periodic. We explicitly construct a Transformer (with linear, exponential, or softmax attention) that learns the mapping f in-context through a causal kernel descent method. The causal kernel descent method we propose provably estimates $x_{t+1} $ based solely on past and current observations $ (x_1, \dots, x_t) $, with connections to the Kaczmarz algorithm in Hilbert spaces. We present experimental results that validate our theoretical findings and suggest their applicability to more general mappings f.

119. Cocoon: Robust Multi-Modal Perception with Uncertainty-Aware Sensor Fusion

链接：https://iclr.cc/virtual/2025/poster/30466 abstract： An important paradigm in 3D object detection is the use of multiple modalities to enhance accuracy in both normal and challenging conditions, particularly for long-tail scenarios. To address this, recent studies have explored two directions of adaptive approaches: MoE-based adaptive fusion, which struggles with uncertainties arising from distinct object configurations, and late fusion for output-level adaptive fusion, which relies on separate detection pipelines and limits comprehensive understanding. In this work, we introduce Cocoon, an object- and feature-level uncertainty-aware fusion framework. The key innovation lies in uncertainty quantification for heterogeneous representations, enabling fair comparison across modalities through the introduction of a feature aligner and a learnable surrogate ground truth, termed feature impression. We also define a training objective to ensure that their relationship provides a valid metric for uncertainty quantification. Cocoon consistently outperforms existing static and adaptive methods in both normal and challenging conditions, including those with natural and artificial corruptions. Furthermore, we show the validity and efficacy of our uncertainty metric across diverse datasets.

120. High-dimensional Analysis of Knowledge Distillation: Weak-to-Strong Generalization and Scaling Laws

链接：https://iclr.cc/virtual/2025/poster/31172 abstract： A growing number of machine learning scenarios rely on knowledge distillation where one uses the output of a surrogate model as labels to supervise the training of a target model. In this work, we provide a sharp characterization of this process for ridgeless, high-dimensional regression, under two settings: (i) model shift, where the surrogate model is arbitrary, and (ii) distribution shift, where the surrogate model is the solution of empirical risk minimization with out-of-distribution data. In both cases, we characterize the precise risk of the target model through non-asymptotic bounds in terms of sample size and data distribution under mild conditions. As a consequence, we identify the form of the optimal surrogate model, which reveals the benefits and limitations of discarding weak features in a data-dependent fashion. In the context of weak-to-strong (W2S) generalization, this has the interpretation that (i) W2S training, with the surrogate as the weak model, can provably outperform training with strong labels under the same data budget, but (ii) it is unable to improve the data scaling law. We validate our results on numerical experiments both on ridgeless regression and on neural network architectures.

121. Learning Task Belief Similarity with Latent Dynamics for Meta-Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/30938 abstract： Meta-reinforcement learning requires utilizing prior task distribution information obtained during exploration to rapidly adapt to unknown tasks. The efficiency of an agent's exploration hinges on accurately identifying the current task. Recent Bayes-Adaptive Deep RL approaches often rely on reconstructing the environment's reward signal, which is challenging in sparse reward settings, leading to suboptimal exploitation. Inspired by bisimulation metrics, which robustly extracts behavioral similarity in continuous MDPs, we propose SimBelief—a novel meta-RL framework via measuring similarity of task belief in Bayes-Adaptive MDP (BAMDP). SimBelief effectively extracts common features of similar task distributions, enabling efficient task identification and exploration in sparse reward environments. We introduce latent task belief metric to learn the common structure of similar tasks and incorporate it into the real task belief. By learning the latent dynamics across task distributions, we connect shared latent task belief features with specific task features, facilitating rapid task identification and adaptation. Our method outperforms state-of-the-art baselines on sparse reward MuJoCo and panda-gym tasks.

122. LongPO: Long Context Self-Evolution of Large Language Models through Short-to-Long Preference Optimization

链接：https://iclr.cc/virtual/2025/poster/28247 abstract： Large Language Models (LLMs) have demonstrated remarkable capabilities through pretraining and alignment. However, superior short-context LLMs may underperform in long-context scenarios due to insufficient long-context alignment. This alignment process remains challenging due to the impracticality of human annotation for extended contexts and the difficulty in balancing short- and long-context performance. To address these

challenges, we introduce LongPO, that enables short-context LLMs to self-evolve to excel on long-context tasks by internally transferring short-context capabilities. LongPO harnesses LLMs to learn from self-generated short-to-long preference data, comprising paired responses generated for identical instructions with long-context inputs and their compressed short-context counterparts, respectively. This preference reveals capabilities and potentials of LLMs cultivated during short-context alignment that may be diminished in under-aligned long-context scenarios. Additionally, LongPO incorporates a short-to-long KL constraint to mitigate short-context performance decline during long-context alignment. When applied to Mistral-7B-Instruct-v0.2 from 128K to 512K context lengths, LongPO fully retains short-context performance and largely outperforms naive SFT and DPO in both long- and short-context tasks. Specifically, LongPO-trained models can achieve results on long-context benchmarks comparable to, or even surpassing, those of superior LLMs (e.g., GPT-4-128K) that involve extensive long-context annotation and larger parameter scales. Our code is available at https://github.com/DAMO-NLP-SG/LongPO.

123. CryoGEN: Generative Energy-based Models for Cryogenic Electron Tomography Reconstruction

链接：https://iclr.cc/virtual/2025/poster/27980 abstract： Cryogenic electron tomography (Cryo-ET) is a powerful technique for visualizing subcellular structures in their native states. Nonetheless, its effectiveness is compromised by anisotropic resolution artifacts caused by the missing-wedge effect. To address this, IsoNet, a deep learning-based method, proposes iteratively reconstructing the missing-wedge information. While successful, IsoNet's dependence on recursive prediction updates often leads to training instability and model divergence. In this study, we introduce CryoGEN—an energy-based probabilistic model that not only mitigates resolution anisotropy but also removes the need for recursive subtomogram averaging, delivering an approximate *10*$\times$ speedup for training. Evaluations across various biological datasets, including immature HIV-1 virions and ribosomes, demonstrate that CryoGEN significantly enhances structural completeness and interpretability of the reconstructed samples.

124. CLIBD: Bridging Vision and Genomics for Biodiversity Monitoring at Scale

链接：https://iclr.cc/virtual/2025/poster/29009 abstract： Measuring biodiversity is crucial for understanding ecosystem health. While prior works have developed machine learning models for taxonomic classification of photographic images and DNA separately, in this work, we introduce a multi-modal approach combining both, using CLIP-style contrastive learning to align images, barcode DNA, and text-based representations of taxonomic labels in a unified embedding space. This allows for accurate classification of both known and unknown insect species without task-specific fine-tuning, leveraging contrastive learning for the first time to fuse DNA and image data. Our method surpasses previous single-modality approaches in accuracy by over 8% on zero-shot learning tasks, showcasing its effectiveness in biodiversity studies.

125. MathCoder2: Better Math Reasoning from Continued Pretraining on Model-translated Mathematical Code

链接：https://iclr.cc/virtual/2025/poster/31210 abstract： Code has been shown to be effective in enhancing the mathematical reasoning abilities of large language models due to its precision and accuracy. Previous works involving continued mathematical pretraining often include code that utilizes math-related packages, which are primarily designed for fields such as engineering, machine learning, signal processing, or module testing, rather than being directly focused on mathematical reasoning. In this paper, we introduce a novel method for generating mathematical code accompanied with corresponding reasoning steps for continued pretraining. Our approach begins with the construction of a high-quality mathematical continued pretraining dataset by incorporating math-related web data, code using mathematical packages, math textbooks, and synthetic data. Next, we construct reasoning steps by extracting LaTeX expressions, the conditions needed for the expressions, and the results of the expressions from the previously collected dataset. Based on this extracted information, we generate corresponding code to accurately capture the mathematical reasoning process. Appending the generated code to each reasoning step results in data consisting of paired natural language reasoning steps and their corresponding code. Combining this data with the original dataset results in a 19.2B-token high-performing mathematical pretraining corpus, which we name MathCode-Pile. Training several popular base models with this corpus significantly improves their mathematical abilities, leading to the creation of the MathCoder2 family of models. All of our data processing and training code is open-sourced, ensuring full transparency and easy reproducibility of the entire data collection and training pipeline.

126. Where Am I and What Will I See: An Auto-Regressive Model for Spatial Localization and View Prediction

链接：https://iclr.cc/virtual/2025/poster/29853 abstract： Spatial intelligence is the ability of a machine to perceive, reason, and act in three dimensions within space and time.Recent advancements in large-scale auto-regressive models have demonstrated remarkable capabilities across various reasoning tasks. However, these models often struggle with fundamental aspects of spatial reasoning, particularly in answering questions like "Where am I?" and "What will I see?". While some attempts have been done, existing approaches typically treat them as separate tasks, failing to capture their interconnected nature. In this paper, we present Generative Spatial Transformer (GST), a novel auto-regressive framework that jointly addresses spatial localization and view prediction. Our model simultaneously estimates the camera pose from a single image and predicts the view from a new

camera pose, effectively bridging the gap between spatial awareness and visual prediction. The proposed innovative camera tokenization method enables the model to learn the joint distribution of 2D projections and their corresponding spatial perspectives in an auto-regressive manner. This unified training paradigm demonstrates that joint optimization of pose estimation and novel view synthesis leads to improved performance in both tasks, for the first time, highlighting the inherent relationship between spatial awareness and visual prediction.

127. KAN: Kolmogorov–Arnold Networks

链接：https://iclr.cc/virtual/2025/poster/29784 abstract： Inspired by the Kolmogorov-Arnold representation theorem, we propose Kolmogorov-Arnold Networks (KANs) as promising alternatives to Multi-Layer Perceptrons (MLPs). While MLPs have fixed activation functions on nodes ("neurons"), KANs have learnable activation functions on edges ("weights"). KANs have no linear weights at all -- every weight parameter is replaced by a univariate function parametrized as a spline. We show that this seemingly simple change makes KANs outperform MLPs in terms of accuracy and interpretability, on small-scale AI + Science tasks. For accuracy, smaller KANs can achieve comparable or better accuracy than larger MLPs in function fitting tasks. Theoretically and empirically, KANs possess faster neural scaling laws than MLPs. For interpretability, KANs can be intuitively visualized and can easily interact with human users. Through two examples in mathematics and physics, KANs are shown to be useful ``collaborators" helping scientists (re)discover mathematical and physical laws. In summary, KANs are promising alternatives for MLPs. Despite the slow training of KANs, their improved accuracy and interpretability show the potential to improve today's deep learning models which rely heavily on MLPs. More research is necessary to make KANs' training more efficient.

128. On the expressiveness and spectral bias of KANs

链接：https://iclr.cc/virtual/2025/poster/27709 abstract： Kolmogorov-Arnold Networks (KAN) \cite{liu2024kan} were very recently proposed as a potential alternative to the prevalent architectural backbone of many deep learning models, the multi-layer perceptron (MLP). KANs have seen success in various tasks of AI for science, with their empirical efficiency and accuracy demonstrated in function regression, PDE solving, and many more scientific problems. In this article, we revisit the comparison of KANs and MLPs, with emphasis on a theoretical perspective. On the one hand, we compare the representation and approximation capabilities of KANs and MLPs. We establish that MLPs can be represented using KANs of a comparable size. This shows that the approximation and representation capabilities of KANs are at least as good as MLPs. Conversely, we show that KANs can be represented using MLPs, but that in this representation the number of parameters increases by a factor of the KAN grid size. This suggests that KANs with a large grid size may be more efficient than MLPs at approximating certain functions. On the other hand, from the perspective of learning and optimization, we study the spectral bias of KANs compared with MLPs. We demonstrate that KANs are less biased toward low frequencies than MLPs. We highlight that the multi-level learning feature specific to KANs, i.e. grid extension of splines, improves the learning process for high-frequency components. Detailed comparisons with different choices of depth, width, and grid sizes of KANs are made, shedding some light on how to choose the hyperparameters in practice.

129. Online Clustering with Nearly Optimal Consistency

链接：https://iclr.cc/virtual/2025/poster/29895 abstract： We give online algorithms for k-Means(more generally, (k, z)-Clustering) with nearly optimal consistency (a notion suggested by Lattanzi & Vassilvitskii (2017)). Our result turns any α-approximate offline algorithm for clustering into an $(1+\epsilon)\alpha^2$-competitive online algorithm for clustering with $O(k \text{poly} \log n)$ consistency. This consistency bound is optimal up to $\text{poly} \log(n)$ factors. Plugging in the offline algorithm that returns the exact optimal solution, we obtain the first$(1 + \epsilon)$-competitive online algorithm for clustering that achieves a linear in k consistency.This simultaneously improves several previous results (Lattanzi & Vassilvitskii, 2017; Fichtenberger et al., 2021). We validate the performance of our algorithm on real datasets by plugging in the practically efficient k-Means++ algorithm. Our online algorithm makes k-Means++ achieve good consistency with little overhead to the quality of solutions.

130. Simplifying, Stabilizing and Scaling Continuous-time Consistency Models

链接：https://iclr.cc/virtual/2025/poster/29963 abstract： Consistency models (CMs) are a powerful class of diffusion-based generative models optimized for fast sampling. Most existing CMs are trained using discretized timesteps, which introduce additional hyperparameters and are prone to discretization errors. While continuous-time formulations can mitigate these issues, their success has been limited by training instability. To address this, we propose a simplified theoretical framework that unifies previous parameterizations of diffusion models and CMs, identifying the root causes of instability. Based on this analysis, we introduce key improvements in diffusion process parameterization, network architecture, and training objectives. These changes enable us to train continuous-time CMs at an unprecedented scale, reaching 1.5B parameters on ImageNet 512×512. Our proposed training algorithm, using only two sampling steps, achieves FID scores of 2.06 on CIFAR-10, 1.48 on ImageNet 64×64, and 1.88 on ImageNet 512×512, narrowing the gap in FID scores with the best existing diffusion models to within 10\%.

131. The Geometry of Categorical and Hierarchical Concepts in Large

Language Models

链接：https://iclr.cc/virtual/2025/poster/29106 abstract： The linear representation hypothesis is the informal idea that semantic concepts are encoded as linear directions in the representation spaces of large language models (LLMs). Previous work has shown how to make this notion precise for representing binary concepts that have natural contrasts (e.g., {male, female}) as directions in representation space. However, many natural concepts do not have natural contrasts (e.g., whether the output is about an animal). In this work, we show how to extend the formalization of the linear representation hypothesis to represent features (e.g., isanimal) as _vectors. This allows us to immediately formalize the representation of categorical concepts as polytopes in the representation space. Further, we use the formalization to prove a relationship between the hierarchical structure of concepts and the geometry of their representations. We validate these theoretical results on the Gemma and LLaMA-3 large language models, estimating representations for 900+ hierarchically related concepts using data from WordNet.

132. Kinetix: Investigating the Training of General Agents through Open-Ended Physics-Based Control Tasks

链接：https://iclr.cc/virtual/2025/poster/27683 abstract： While large models trained with self-supervised learning on offline datasets have shown remarkable capabilities in text and image domains, achieving the same generalisation for agents that act in sequential decision problems remains an open challenge.In this work, we take a step towards this goal by procedurally generating tens of millions of 2D physics-based tasks and using these to train a general reinforcement learning (RL) agent for physical control.To this end, we introduce Kinetix: an open-ended space of physics-based RL environments that can represent tasks ranging from robotic locomotion and grasping to video games and classic RL environments, all within a unified framework.Kinetix makes use of our novel hardware-accelerated physics engine Jax2D that allows us to cheaply simulate billions of environment steps during training.Our trained agent exhibits strong physical reasoning capabilities in 2D space, being able to zero-shot solve unseen human-designed environments. Furthermore, fine-tuning this general agent on tasks of interest shows significantly stronger performance than training an RL agent tabula rasa. This includes solving some environments that standard RL training completely fails at.We believe this demonstrates the feasibility of large scale, mixed-quality pre-training for online RL and we hope that Kinetix will serve as a useful framework to investigate this further.

133. Accelerating neural network training: An analysis of the AlgoPerf competition

链接：https://iclr.cc/virtual/2025/poster/30490 abstract： The goal of the AlgoPerf: Training Algorithms competition is to evaluate practical speed-ups in neural network training achieved solely by improving the underlying training algorithms. In the external tuning ruleset, submissions must provide workload-agnostic hyperparameter search spaces, while in the self-tuning ruleset they must be completely hyperparameter-free. In both rulesets, submissions are compared on time-to-result across multiple deep learning workloads, training on fixed hardware. This paper presents the inaugural AlgoPerf competition's results, which drew 18 diverse submissions from 10 teams. Our investigation reveals several key findings: (1) The winning submission in the external tuning ruleset, using Distributed Shampoo, demonstrates the effectiveness of non-diagonal preconditioning over popular methods like Adam, even when compared on wall-clock runtime. (2) The winning submission in the self-tuning ruleset, based on the Schedule Free AdamW algorithm, demonstrates a new level of effectiveness for completely hyperparameter-free training algorithms. (3) The top-scoring submissions were surprisingly robust to workload changes. We also discuss the engineering challenges encountered in ensuring a fair comparison between different training algorithms. These results highlight both the significant progress so far, and the considerable room for further improvements.

134. An Auditing Test to Detect Behavioral Shift in Language Models

链接：https://iclr.cc/virtual/2025/poster/28786 abstract： As language models (LMs) approach human-level performance, a comprehensive understanding of their behavior becomes crucial. This includes evaluating capabilities, biases, task performance, and alignment with societal values. Extensive initial evaluations, including red teaming and diverse benchmarking, can establish a model’s behavioral profile. However, subsequent fine-tuning or deployment modifications may alter these behaviors in unintended ways. We present an efficient statistical test to tackle Behavioral Shift Auditing (BSA) in LMs, which we define as detecting distribution shifts in qualitative properties of the output distributions of LMs. Our test compares model generations from a baseline model to those of the model under scrutiny and provides theoretical guarantees for change detection while controlling false positives. The test features a configurable tolerance parameter that adjusts sensitivity to behavioral changes for different use cases. We evaluate our approach using two case studies: monitoring changes in (a) toxicity and (b) translation performance. We find that the test is able to detect meaningful changes in behavior distributions using just hundreds of examples.

135. Adversarial Training Can Provably Improve Robustness: Theoretical Analysis of Feature Learning Process Under Structured Data

链接：https://iclr.cc/virtual/2025/poster/28674 abstract： Adversarial training is a widely-applied approach to training deep neural networks to be robust against adversarial perturbation. However, although adversarial training has achieved empirical success in practice, it still remains unclear why adversarial examples exist and how adversarial training methods improve model

robustness. In this paper, we provide a theoretical understanding of adversarial examples and adversarial training algorithms from the perspective of feature learning theory. Specifically, we focus on a multiple classification setting, where the structured data can be composed of two types of features: the robust features, which are resistant to perturbation but sparse, and the non-robust features, which are susceptible to perturbation but dense. We train a two-layer smoothed ReLU convolutional neural network to learn our structured data. First, we prove that by using standard training (gradient descent over the empirical risk), the network learner primarily learns the non-robust feature rather than the robust feature, which thereby leads to the adversarial examples that are generated by perturbations aligned with negative non-robust feature directions. Then, we consider the gradient-based adversarial training algorithm, which runs gradient ascent to find adversarial examples and runs gradient descent over the empirical risk at adversarial examples to update models. We show that the adversarial training method can provably strengthen the robust feature learning and suppress the non-robust feature learning to improve the network robustness. Finally, we also empirically validate our theoretical findings with experiments on real-image datasets, including MNIST, CIFAR10 and SVHN.

136. TRENDy: Temporal Regression of Effective Nonlinear Dynamics

链接：https://iclr.cc/virtual/2025/poster/29850 abstract： Spatiotemporal dynamics pervade the natural sciences, from the morphogen dynamics underlying patterning in animal pigmentation to the protein waves controlling cell division. A central challenge lies in understanding how controllable parameters induce qualitative changes in system behavior called bifurcations. This endeavor is particularly difficult in realistic settings where governing partial differential equations (PDEs) are unknown and data is limited and noisy. To address this challenge, we propose TRENDy (Temporal Regression of Effective Nonlinear Dynamics), an equation-free approach to learning low-dimensional, predictive models of spatiotemporal dynamics. TRENDy first maps input data to a low-dimensional space of effective dynamics through a cascade of multiscale filtering operations. Our key insight is the recognition that these effective dynamics can be fit by a neural ordinary differential equation (NODE) having the same parameter space as the input PDE. The preceding filtering operations strongly regularize the phase space of the NODE, making TRENDy significantly more robust to noise compared to existing methods. We train TRENDy to predict the effective dynamics of synthetic and real data representing dynamics from across the physical and life sciences. We then demonstrate how we can automatically locate both Turing and Hopf bifurcations in unseen regions of parameter space. We finally apply our method to the analysis of spatial patterning of the ocellated lizard through development. We found that TRENDy's predicted effective state not only accurately predicts spatial changes over time but also identifies distinct pattern features unique to different anatomical regions, such as the tail, neck, and body--an insight that highlights the potential influence of surface geometry on reaction-diffusion mechanisms and their role in driving spatially varying pattern dynamics.

137. HQGS: High-Quality Novel View Synthesis with Gaussian Splatting in Degraded Scenes

链接：https://iclr.cc/virtual/2025/poster/31164 abstract： 3D Gaussian Splatting (3DGS) has shown promising results for Novel View Synthesis. However, while it is quite effective when based on high-quality images, its performance declines as image quality degrades, due to lack of resolution, motion blur, noise, compression artifacts, or other factors common in real-world data collection. While some solutions have been proposed for specific types of degradation, general techniques are still missing. To address the problem, we propose a robust HQGS that significantly enhances the 3DGS under various degradation scenarios. We first analyze that 3DGS lacks sufficient attention in some detailed regions in low-quality scenes, leading to the absence of Gaussian primitives in those areas and resulting in loss of detail in the rendered images. To address this issue, we focus on leveraging edge structural information to provide additional guidance for 3DGS, enhancing its robustness. First, we introduce an edge-semantic fusion guidance module that combines rich texture information from high-frequency edge-aware maps with semantic information from images. The fused features serve as prior guidance to capture detailed distribution across different regions, bringing more attention to areas with detailed edge information and allowing for a higher concentration of Gaussian primitives to be assigned to such areas. Additionally, we present a structural cosine similarity loss to complement pixel-level constraints, further improving the quality of the rendered images. Extensive experiments demonstrate that our method offers better robustness and achieves the best results across various degraded scenes. Source code and trained models are publicly available at: \url{https://github.com/linxin0/HQGS}.

138. Dynamic Contrastive Skill Learning with State-Transition Based Skill Clustering and Dynamic Length Adjustment

链接：https://iclr.cc/virtual/2025/poster/30754 abstract： Reinforcement learning (RL) has made significant progress in various domains, but scaling it to long-horizon tasks with complex decision-making remains challenging. Skill learning attempts to address this by abstracting actions into higher-level behaviors. However, current approaches often fail to recognize semantically similar behaviors as the same skill and use fixed skill lengths, limiting flexibility and generalization. To address this, we propose Dynamic Contrastive Skill Learning (DCSL), a novel framework that redefines skill representation and learning. DCSL introduces three key ideas: state-transition based skill definition, skill similarity function learning, and dynamic skill length adjustment. By focusing on state transitions and leveraging contrastive learning, DCSL effectively captures the semantic context of behaviors and adapts skill lengths to match the appropriate temporal extent of behaviors. Our approach enables more flexible and adaptive skill extraction, particularly in complex or noisy datasets, and demonstrates competitive performance compared to existing methods in task completion and efficiency.

139. Efficient Causal Decision Making with One-sided Feedback

链接：https://iclr.cc/virtual/2025/poster/29469 abstract： We study a class of decision-making problems with one-sided feedback, where outcomes are only observable for specific actions. A typical example is bank loans, where the repayment status is known only if a loan is approved and remains undefined if rejected. In such scenarios, conventional approaches to causal decision evaluation and learning from observational data are not directly applicable. In this paper, we introduce a novel value function to evaluate decision rules that addresses the issue of undefined counterfactual outcomes. Without assuming no unmeasured confounders, we establish the identification of the value function using shadow variables. Furthermore, leveraging semiparametric theory, we derive the efficiency bound for the proposed value function and develop efficient methods for decision evaluation and learning. Numerical experiments and a real-world data application demonstrate the empirical performance of our proposed methods.

140. Think Then React: Towards Unconstrained Action-to-Reaction Motion Generation

链接：https://iclr.cc/virtual/2025/poster/29437 abstract： Modeling human-like action-to-reaction generation has significant real-world applications, like human-robot interaction and games.Despite recent advancements in single-person motion generation, it is still challenging to well handle action-to-reaction generation, due to the difficulty of directly predicting reaction from action sequence without prompts, and the absence of a unified representation that effectively encodes multi-person motion.To address these challenges, we introduce Think-Then-React (TTR), a large language-model-based framework designed to generate human-like reactions.First, with our fine-grained multimodal training strategy, TTR is capable to unify two processes during inference: a thinking process that explicitly infers action intentions and reasons corresponding reaction description, which serve as semantic prompts, and a reacting process that predicts reactions based on input action and the inferred semantic prompts.Second, to effectively represent multi-person motion in language models, we propose a unified motion tokenizer by decoupling egocentric pose and absolute space features, which effectively represents action and reaction motion with same encoding.Extensive experiments demonstrate that TTR outperforms existing baselines, achieving significant improvements in evaluation metrics, such as reducing FID from 3.988 to 1.942.

141. Ferret-UI 2: Mastering Universal User Interface Understanding Across Platforms

链接：https://iclr.cc/virtual/2025/poster/32099 abstract： Building a generalist model for user interface (UI) understanding is challenging due to various foundational issues, such as platform diversity, resolution variation, and data limitation. In this paper, we introduce Ferret-UI 2, a multimodal large language model (MLLM) designed for universal UI understanding across a wide range of platforms, including iPhone, Android, iPad, Webpage, and AppleTV. Building on the foundation of Ferret-UI, Ferret-UI 2 introduces three key innovations: support for multiple platform types, high-resolution perception through adaptive scaling, and advanced task training data generation powered by GPT-4o with set-of-mark visual prompting. These advancements enable Ferret-UI 2 to perform complex, user-centered interactions, making it highly versatile and adaptable for the expanding diversity of platform ecosystems. Extensive empirical experiments on referring, grounding, user-centric advanced tasks (comprising 9 subtasks $\times$ 5 platforms), GUIDE next-action prediction dataset, and GUI-World multi-platform benchmark demonstrate that Ferret-UI 2 significantly outperforms Ferret-UI, and also shows strong cross-platform transfer capabilities.

142. Regularized Proportional Fairness Mechanism for Resource Allocation Without Money

链接：https://iclr.cc/virtual/2025/poster/31454 abstract： Mechanism design in resource allocation studies dividing limited resources among self-interested agents whose satisfaction with the allocation depends on privately held utilities. We consider the problem in a payment-free setting, with the aim of maximizing social welfare while enforcing incentive compatibility (IC), i.e., agents cannot inflate allocations by misreporting their utilities. The well-known proportional fairness (PF) mechanism achieves the maximum possible social welfare but incurs an undesirably high exploitability (the maximum unilateral inflation in utility from misreport and a measure of deviation from IC). In fact, it is known that no mechanism can achieve the maximum social welfare and exact incentive compatibility (IC) simultaneously without the use of monetary incentives (Cole et al., 2013). Motivated by this fact, we propose learning an approximate mechanism that desirably trades off the competing objectives. Our main contribution is to design an innovative neural network architecture tailored to the resource allocation problem, which we name Regularized Proportional Fairness Network (RPF-Net). RPF-Net regularizes the output of the PF mechanism by a learned function approximator of the most exploitable allocation, with the aim of reducing the incentive for any agent to misreport. We derive generalization bounds that guarantee the mechanism performance when trained under finite and out-of-distribution samples and experimentally demonstrate the merits of the proposed mechanism compared to the state-of-the-art. The PF mechanism acts as an important benchmark for comparing the social welfare of any mechanism. However, there exists no established way of computing its exploitability. The challenge here is that we need to find the maximizer of an optimization problem for which the gradient is only implicitly defined. We for the first time provide a systematic method for finding such (sub)gradients, which enables the evaluation of the exploitability of the PF mechanism through iterative (sub)gradient ascent.

143. Dynamic Neural Fortresses: An Adaptive Shield for Model Extraction

Defense

链接：https://iclr.cc/virtual/2025/poster/31275 abstract： Model extraction aims to acquire a pre-trained black-box model concealed behind a black-box API. Existing defense strategies against model extraction primarily concentrate on preventing the unauthorized extraction of API functionality. However, two significant challenges still need to be solved: (i) Neural network architecture of the API constitutes a form of intellectual property that also requires protection; (ii) The current practice of allocating the same network architecture to both attack and benign queries results in substantial resource wastage. To address these challenges, we propose a novel \textit{Dynamic Neural Fortresses} (DNF) defense method, employing a dynamic Early-Exit neural network, deviating from the conventional fixed architecture. Firstly, we facilitate the random exit of attack queries from the network at earlier layers. This strategic exit point selection significantly reduces the computational cost for attack queries. Furthermore, the random exit of attack queries from earlier layers introduces increased uncertainty for attackers attempting to discern the exact architecture, thereby enhancing architectural protection. On the contrary, we aim to facilitate benign queries to exit at later layers, preserving model utility, as these layers typically yield meaningful information. Extensive experiments on defending against various model extraction scenarios and datasets demonstrate the effectiveness of DNF, achieving a notable 2$\times$ improvement in efficiency and an impressive reduction of up to 12\% in clone model accuracy compared to SOTA defense methods. Additionally, DNF provides strong protection against neural architecture theft, effectively safeguarding network architecture from being stolen.

144. Agent S: An Open Agentic Framework that Uses Computers Like a Human

链接：https://iclr.cc/virtual/2025/poster/28525 abstract： We present Agent S, an open agentic framework that enables autonomous interaction with computers through Graphical User Interface (GUI), aimed at transforming human-computer interaction by automating complex, multi-step tasks. Agent S addresses three key challenges in automating computer tasks: acquiring domain-specific knowledge, planning over long task horizons, and handling dynamic, non-uniform interfaces. To this end, Agent S introduces experience-augmented hierarchical planning, which learns from external knowledge search and internal experience retrieval at multiple levels, facilitating efficient task planning and subtask execution. In addition, it employs an Agent-Computer Interface (ACI) to better elicit the reasoning and control capabilities of GUI agents based on Multimodal Large Language Models (MLLMs). Evaluation on the OSWorld benchmark shows that Agent S outperforms the baseline by 9.37\% on success rate (an 83.6\% relative improvement) and achieves a new state-of-the-art. Comprehensive analysis highlights the effectiveness of individual components and provides insights for future improvements. Furthermore, Agent S demonstrates broad generalizability to different operating systems on a newly-released WindowsAgentArena benchmark. Code available at https://github.com/simular-ai/Agent-S.

145. Simple ReFlow: Improved Techniques for Fast Flow Models

链接：https://iclr.cc/virtual/2025/poster/28855 abstract： Diffusion and flow-matching models achieve remarkable generative performance but at the cost of many neural function evaluations (NFE), which slows inference and limits applicability to time-critical tasks. The ReFlow procedure can accelerate sampling by straightening generation trajectories. But it is an iterative procedure, typically requiring training on simulated data, and results in reduced sample quality. To mitigate sample deterioration, we examine the design space of ReFlow and highlight potential pitfalls in prior heuristic practices. We then propose seven improvements for training dynamics, learning and inference, which are verified with thorough ablation studies on CIFAR10 32×32, AFHQv2 64×64, and FFHQ 64×64. Combining all our techniques, we achieve state-of-the-art FID scores (without / with guidance, resp.) for fast generation via neural ODEs: 2.23 / 1.98 on CIFAR10, 2.30 / 1.91 on AFHQv2, 2.84 / 2.67 on FFHQ, and 3.49 / 1.74 on ImageNet-64, all with merely 9 NFEs.

146. Protein Language Model Fitness is a Matter of Preference

链接：https://iclr.cc/virtual/2025/poster/29443 abstract： Leveraging billions of years of evolution, scientists have trained protein language models (pLMs) to understand the sequence and structure space of proteins aiding in the design of more functional proteins. Although they have shown ability to improve efficiency in engineering, it remains unclear under what conditions they will succeed or fail. We aim to predict the circumstances in which pLMs can successfully perform zero-shot fitness estimation. Our work demonstrates the trends observed over hundreds of deep mutational scans across multiple different fitness objectives. We find that the likelihood, or abstractly, implicit preference of a certain protein sequence imbued during pretraining is predictive fitness prediction capabilities. Both over-preferred and under-preferred wild type sequences harm performance. Generating a causal link between training data and likelihood, we show a power law tail over what data increases protein likelihood which is tied to training sequence homology. Lastly, proteins of low likelihood can be remedied by unsupervised finetuning. In sum, the zero-shot fitness estimation abilities of pLMs can be predicted by the likelihood of the engineered sequence, thus suggesting when pLMs should be deployed in protein maturation campaigns and a way to improve their performance under circumstances of low likelihood.

147. Forte : Finding Outliers with Representation Typicality Estimation

链接：https://iclr.cc/virtual/2025/poster/30817 abstract： Generative models can now produce photorealistic synthetic data which is virtually indistinguishable from the real data used to train it. This is a significant evolution over previous models which

could produce reasonable facsimiles of the training data, but ones which could be visually distinguished from the training data by human evaluation. Recent work on OOD detection has raised doubts that generative model likelihoods are optimal OOD detectors due to issues involving likelihood misestimation, entropy in the generative process, and typicality. We speculate that generative OOD detectors also failed because their models focused on the pixels rather than the semantic content of the data, leading to failures in near-OOD cases where the pixels may be similar but the information content is significantly different. We hypothesize that estimating typical sets using self-supervised learners leads to better OOD detectors. We introduce a novel approach that leverages representation learning, and informative summary statistics based on manifold estimation, to address all of the aforementioned issues. Our method outperforms other unsupervised approaches and achieves state-of-the art performance on well-established challenging benchmarks, and new synthetic data detection tasks.

148. Intelligent Go-Explore: Standing on the Shoulders of Giant Foundation Models

链接：https://iclr.cc/virtual/2025/poster/29147 abstract： Go-Explore is a powerful family of algorithms designed to solve hard-exploration problems built on the principle of archiving discovered states, and iteratively returning to and exploring from the most promising states. This approach has led to superhuman performance across a wide variety of challenging problems including Atari games and robotic control, but requires manually designing heuristics to guide exploration (i.e., determine which states to save and explore from, and what actions to consider next), which is time-consuming and infeasible in general. To resolve this, we propose Intelligent Go-Explore (IGE) which greatly extends the scope of the original Go-Explore by replacing these handcrafted heuristics with the intelligence and internalized human notions of interestingness captured by giant pretrained foundation models (FMs). This provides IGE with a human-like ability to instinctively identify how interesting or promising any new state is (e.g., discovering new objects, locations, or behaviors), even in complex environments where heuristics are hard to define. Moreover, IGE offers the exciting opportunity to recognize and capitalize on serendipitous discoveries---states encountered during exploration that are valuable in terms of exploration, yet where what makes them interesting was not anticipated by the human user. We evaluate our algorithm on a diverse range of language and vision-based tasks that require search and exploration. Across these tasks, IGE strongly exceeds classic reinforcement learning and graph search baselines, and also succeeds where prior state-of-the-art FM agents like Reflexion completely fail. Overall, Intelligent Go-Explore combines the tremendous strengths of FMs and the powerful Go-Explore algorithm, opening up a new frontier of research into creating more generally capable agents with impressive exploration capabilities. All our code is open-sourced at: https://github.com/conglu1997/intelligent-go-explore.

149. MOOSE-Chem: Large Language Models for Rediscovering Unseen Chemistry Scientific Hypotheses

链接：https://iclr.cc/virtual/2025/poster/29319 abstract： Scientific discovery contributes largely to the prosperity of human society, and recent progress shows that LLMs could potentially catalyst the process. However, it is still unclear whether LLMs can discover novel and valid hypotheses in chemistry. In this work, we investigate this main research question: whether LLMs can automatically discover novel and valid chemistry research hypotheses, given only a research question? With extensive discussions with chemistry experts, we adopt the assumption that a majority of chemistry hypotheses can be resulted from a research background question and several inspirations. With this key insight, we break the main question into three smaller fundamental questions. In brief, they are: (1) given a background question, whether LLMs can retrieve good inspirations; (2) with background and inspirations, whether LLMs can lead to hypothesis; and (3) whether LLMs can identify good hypotheses to rank them higher. To investigate these questions, we construct a benchmark consisting of 51 chemistry papers published in Nature or a similar level in 2024 (all papers are only available online since 2024). Every paper is divided by chemistry PhD students into three components: background, inspirations, and hypothesis. The goal is to rediscover the hypothesis given only the background and a large chemistry literature corpus consisting the ground truth inspiration papers, with LLMs trained with data up to 2023. We also develop an LLM-based multi-agent framework that leverages the assumption, consisting of three stages reflecting the more smaller questions. The proposed method can rediscover many hypotheses with very high similarity with the ground truth ones, covering the main innovations.

150. Automated Design of Agentic Systems

链接：https://iclr.cc/virtual/2025/poster/28073 abstract： Researchers are investing substantial effort in developing powerful general-purpose agents, wherein Foundation Models are used as modules within agentic systems (e.g. Chain-of-Thought, Self-Reflection, Toolformer). However, the history of machine learning teaches us that hand-designed solutions are eventually replaced by learned solutions. We describe a newly forming research area, Automated Design of Agentic Systems (ADAS), which aims to automatically create powerful agentic system designs, including inventing novel building blocks and/or combining them in new ways. We further demonstrate that there is an unexplored yet promising approach within ADAS where agents can be defined in code and new agents can be automatically discovered by a meta agent programming ever better ones in code. Given that programming languages are Turing Complete, this approach theoretically enables the learning of any possible agentic system: including novel prompts, tool use, workflows, and combinations thereof. We present a simple yet effective algorithm named Meta Agent Search to demonstrate this idea, where a meta agent iteratively programs interesting new agents based on an ever-growing archive of previous discoveries. Through extensive experiments across multiple domains including coding, science, and math, we show that our algorithm can progressively invent agents with novel designs that greatly outperform state-of-the-art hand-designed agents. Importantly, we consistently observe the surprising result that agents invented by Meta Agent Search maintain superior performance even when transferred across domains and models, demonstrating their

robustness and generality. Provided we develop it safely, our work illustrates the potential of an exciting new research direction toward automatically designing ever-more powerful agentic systems to benefit humanity.

151. Watch Less, Do More: Implicit Skill Discovery for Video-Conditioned Policy

链接：https://iclr.cc/virtual/2025/poster/28751 abstract： In this paper, we study the problem of video-conditioned policy learning. While previous works mostly focus on learning policies that perform a single skill specified by the given video, we take a step further and aim to learn a policy that can perform multiple skills according to the given video, and generalize to unseen videos by recombining these skills. To solve this problem, we propose our algorithm, Watch-Less-Do-More, an information bottleneck-based imitation learning framework for implicit skill discovery and video-conditioned policy learning. In our method, an information bottleneck objective is employed to control the information contained in the video representation, ensuring that it only encodes information relevant to the current skill (Watch-Less). By discovering potential skills from training videos, the learned policy is able to recombine them and generalize to unseen videos to achieve compositional generalization (Do-More). To evaluate our method, we perform extensive experiments in various environments and show that our algorithm substantially outperforms baselines (up to 2x) in terms of compositional generalization ability.

152. Learning and aligning single-neuron invariance manifolds in visual cortex

链接：https://iclr.cc/virtual/2025/poster/28569 abstract： Understanding how sensory neurons exhibit selectivity to certain features and invariance to others is central to uncovering the computational principles underlying robustness and generalization in visual perception. Most existing methods for characterizing selectivity and invariance identify single or finite discrete sets of stimuli. Since these are only isolated measurements from an underlying continuous manifold, characterizing invariance properties accurately and comparing them across neurons with varying receptive field size, position, and orientation, becomes challenging. Consequently, a systematic analysis of invariance types at the population level remains under-explored. Building on recent advances in learning continuous invariance manifolds, we introduce a novel method to accurately identify and align invariance manifolds of visual sensory neurons, overcoming these challenges. Our approach first learns the continuous invariance manifold of stimuli that maximally excite a neuron modeled by a response-predicting deep neural network. It then learns an affine transformation on the pixel coordinates such that the same manifold activates another neuron as strongly as possible, effectively aligning their invariance manifolds spatially. This alignment provides a principled way to quantify and compare neuronal invariances irrespective of receptive field differences. Using simulated neurons, we demonstrate that our method accurately learns and aligns known invariance manifolds, robustly identifying functional clusters. When applied to macaque V1 neurons, it reveals functional clusters of neurons, including simple and complex cells. Overall, our method enables systematic, quantitative exploration of the neural invariance landscape, to gain new insights into the functional properties of visual sensory neurons.

153. Visual-O1: Understanding Ambiguous Instructions via Multi-modal Multi-turn Chain-of-thoughts Reasoning

链接：https://iclr.cc/virtual/2025/poster/27930 abstract： As large-scale models evolve, language instructions are increasingly utilized in multi-modal tasks. Due to human language habits, these instructions often contain ambiguities in real-world scenarios, necessitating the integration of visual context or common sense for accurate interpretation. However, even highly intelligent large models exhibit observable performance limitations on ambiguous instructions, where weak reasoning abilities of disambiguation can lead to catastrophic errors. To address this issue, this paper proposes Visual-O1, a multi-modal multi-turn chain-of-thought reasoning framework. It simulates human multi-modal multi-turn reasoning, providing instantial experience for highly intelligent models or empirical experience for generally intelligent models to understand ambiguous instructions. Unlike traditional methods that require models to possess high intelligence to understand long texts or perform lengthy complex reasoning, our framework does not notably increase computational overhead and is more general and effective, even for generally intelligent models. Experiments show that our method not only enhances the performance of models of different intelligence levels on ambiguous instructions but also improves their performance on general datasets. Our work highlights the potential of artificial intelligence to work like humans in real-world scenarios with uncertainty and ambiguity. We release our data and code at https://github.com/kodenii/Visual-O1.

154. Learning Video-Conditioned Policy on Unlabelled Data with Joint Embedding Predictive Transformer

链接：https://iclr.cc/virtual/2025/poster/29514 abstract： The video-conditioned policy takes prompt videos of the desired tasks as a condition and is regarded for its prospective generalizability. Despite its promise, training a video-conditioned policy is non-trivial due to the need for abundant demonstrations. In some tasks, the expert rollouts are merely available as videos, and costly and time-consuming efforts are required to annotate action labels. To address this, we explore training video-conditioned policy on a mixture of demonstrations and unlabeled expert videos to reduce reliance on extensive manual annotation. We introduce the Joint Embedding Predictive Transformer (JEPT) to learn a video-conditioned policy through sequence modeling. JEPT is designed to jointly learn visual transition prediction and inverse dynamics. The visual transition is captured from both

demonstrations and expert videos, on the basis of which the inverse dynamics learned from demonstrations is generalizable to the tasks without action labels. Experiments on a series of simulated visual control tasks evaluate that JEPT can effectively leverage the mixture dataset to learn a generalizable policy. JEPT outperforms baselines in the tasks without action-labeled data and unseen tasks. We also experimentally reveal the potential of JEPT as a simple visual priors injection approach to enhance the video-conditioned policy.

155. Discrete Latent Plans via Semantic Skill Abstractions

链接：https://iclr.cc/virtual/2025/poster/30017 abstract： Skill learning from language instructions is a critical challenge in developing intelligent agents that can generalize across diverse tasks and follow complex human instructions. Hierarchical methods address this by decomposing the learning problem into multiple levels, where the high-level and low-level policies are mediated through a latent plan space. Effective modeling and learning of this latent plan space are key to enabling robust and interpretable skill learning. In this paper, we introduce LADS, a hierarchical approach that learns language-conditioned discrete latent plans through semantic skill abstractions. Our method decouples the learning of the latent plan space from the language-conditioned high-level policy to improve training stability. First, we incorporate a trajectory encoder to learn a discrete latent space with the low-level policy, regularized by language instructions. Next, we model the high-level policy as a categorical distribution over these discrete latent plans to capture the multi-modality of the dataset. Through experiments in simulated control environments, we demonstrate that LADS outperforms state-of-the-art methods in both skill learning and compositional generalization.

156. Efficient Residual Learning with Mixture-of-Experts for Universal Dexterous Grasping

链接：https://iclr.cc/virtual/2025/poster/30569 abstract： Universal dexterous grasping across diverse objects presents a fundamental yet formidable challenge in robot learning. Existing approaches using reinforcement learning (RL) to develop policies on extensive object datasets face critical limitations, including complex curriculum design for multi-task learning and limited generalization to unseen objects. To overcome these challenges, we introduce ResDex, a novel approach that integrates residual policy learning with a mixture-of-experts (MoE) framework. ResDex is distinguished by its use of geometry-agnostic base policies that are efficiently acquired on individual objects and capable of generalizing across a wide range of unseen objects. Our MoE framework incorporates several base policies to facilitate diverse grasping styles suitable for various objects. By learning residual actions alongside weights that combine these base policies, ResDex enables efficient multi-task RL for universal dexterous grasping.ResDex achieves state-of-the-art performance on the DexGraspNet dataset comprising 3,200 objects with an 88.8% success rate. It exhibits no generalization gap with unseen objects and demonstrates superior training efficiency, mastering all tasks within only 12 hours on a single GPU. For further details and videos, visit our project page.

157. Adaptive Energy Alignment for Accelerating Test-Time Adaptation

链接：https://iclr.cc/virtual/2025/poster/28139 abstract： In response to the increasing demand for tackling out-of-domain (OOD) scenarios, test-time adaptation (TTA) has garnered significant research attention in recent years. To adapt a source pre-trained model to target samples without getting access to their labels, existing approaches have typically employed entropy minimization (EM) loss as a primary objective function. In this paper, we propose an adaptive energy alignment (AEA) solution that achieves fast online TTA. We start from the re-interpretation of the EM loss by decomposing it into two energy-based terms with conflicting roles, showing that the EM loss can potentially hinder the assertive model adaptation. Our AEA addresses this challenge by strategically reducing the energy gap between the source and target domains during TTA, aiming to effectively align the target domain with the source domains and thus to accelerate adaptation. We specifically propose two novel strategies, each contributing a necessary component for TTA: (i) aligning the energy level of each target sample with the energy zone of the source domain that the pre-trained model is already familiar with, and (ii) precisely guiding the direction of the energy alignment by matching the class-wise correlations between the source and target domains. Our approach demonstrates its effectiveness on various domain shift datasets including CIFAR10-C, CIFAR100-C, and TinyImageNet-C.

158. MLLM as Retriever: Interactively Learning Multimodal Retrieval for Embodied Agents

链接：https://iclr.cc/virtual/2025/poster/30075 abstract： MLLM agents demonstrate potential for complex embodied tasks by retrieving multimodal task-relevant trajectory data. However, current retrieval methods primarily focus on surface-level similarities of textual or visual cues in trajectories, neglecting their effectiveness for the specific task at hand. To address this issue, we propose a novel method, MART, which enhances the performance of embodied agents by utilizing interaction data to fine-tune an MLLM retriever based on preference learning, such that the retriever fully considers the effectiveness of trajectories and prioritize them for unseen tasks. We also introduce Trajectory Abstraction, a mechanism that leverages MLLMs' summarization capabilities to represent trajectories with fewer tokens while preserving key information, enabling agents to better comprehend milestones in the trajectory. Experimental results across various environments demonstrate our method significantly improves task success rates in unseen scenes compared to baseline methods. This work presents a new paradigm for multimodal retrieval in embodied agents, by fine-tuning a general-purpose MLLM as the retriever to assess trajectory effectiveness. All the code for benchmark tasks, simulator modifications and the MLLM retriever is available at https://github.com/PKU-RL/MART.

159. Cross-Embodiment Dexterous Grasping with Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28010 abstract： Dexterous hands exhibit significant potential for complex real-world grasping tasks. While recent studies have primarily focused on learning policies for specific robotic hands, the development of a universal policy that controls diverse dexterous hands remains largely unexplored.In this work, we study the learning of cross-embodiment dexterous grasping policies using reinforcement learning (RL). Inspired by the capability of human hands to control various dexterous hands through teleoperation, we propose a universal action space based on the human hand's eigengrasps. The policy outputs eigengrasp actions that are then converted into specific joint actions for each robot hand through a retargeting mapping. We simplify the robot hand's proprioception to include only the positions of fingertips and the palm, offering a unified observation space across different robot hands. Our approach demonstrates an 80\% success rate in grasping objects from the YCB dataset across four distinct embodiments using a single vision-based policy. Additionally, our policy exhibits zero-shot generalization to two previously unseen embodiments and significant improvement in efficient finetuning. For further details and videos, visit our project page (https://sites.google.com/view/crossdex).

160. From Pixels to Tokens: Byte-Pair Encoding on Quantized Visual Modalities

链接：https://iclr.cc/virtual/2025/poster/31074 abstract： Multimodal Large Language Models have made significant strides in integrating visual and textual information, yet they often struggle with effectively aligning these modalities. We introduce a novel image tokenizer that bridges this gap by applying the principle of Byte-Pair Encoding (BPE) to visual data. Unlike conventional approaches that rely on separate visual encoders, our method directly incorporates structural prior information into image tokens, mirroring the successful tokenization strategies used in text-only Large Language Models. This innovative approach enables Transformer models to more effectively learn and reason across modalities. Through theoretical analysis and extensive experiments, we demonstrate that our BPE Image Tokenizer significantly enhances MLLMs' multimodal understanding capabilities, even with limited training data. Leveraging this method, we develop Being-VL-0, a model that demonstrates superior performance across various benchmarks and shows promising scalability, potentially paving the way for more efficient and capable multimodal foundation models. For further details, visit our website https://github.com/BeingBeyond/Being-VL-0.

161. Knowing Your Target: Target-Aware Transformer Makes Better Spatio-Temporal Video Grounding

链接：https://iclr.cc/virtual/2025/poster/29364 abstract： Transformer has attracted increasing interest in spatio-temporal video grounding, or STVG, owing to its end-to-end pipeline and promising result. Existing Transformer-based STVG approaches often leverage a set of object queries, which are initialized simply using zeros and then gradually learn target position information via iterative interactions with multimodal features, for spatial and temporal localization. Despite simplicity, these zero object queries, due to lacking target-specific cues, are hard to learn discriminative target information from interactions with multimodal features in complicated scenarios (e.g., with distractors or occlusion), resulting in degradation. Addressing this, we introduce a novel $\textbf{T}$arget-$\textbf{A}$ware Transformer for $\textbf{STVG}$ ($\textbf{TA-STVG}$), which seeks to adaptively generate object queries via exploring target-specific cues from the given video-text pair, for improving STVG. The key lies in two simple yet effective modules, comprising text-guided temporal sampling (TTS) and attribute-aware spatial activation (ASA), working in a cascade. The former focuses on selecting target-relevant temporal cues from a video utilizing holistic text information, while the latter aims at further exploiting the fine-grained visual attribute information of the object from previous target-aware temporal cues, which is applied for object query initialization. Compared to existing methods leveraging zero-initialized queries, object queries in our TA-STVG, directly generated from a given video-text pair, naturally carry target-specific cues, making them adaptive and better interact with multimodal features for learning more discriminative information to improve STVG. In our experiments on three benchmarks, including HCSTVG-v1/-v2 and VidSTG, TA-STVG achieves state-of-the-art performance and significantly outperforms the baseline, validating its efficacy. Moreover, TTS and ASA are designed for general purpose. When applied to existing methods such as TubeDETR and STCAT, we show substantial performance gains, verifying its generality. Code is released at https://github.com/HengLan/TA-STVG.

162. Cross-Domain Offline Policy Adaptation with Optimal Transport and Dataset Constraint

链接：https://iclr.cc/virtual/2025/poster/29997 abstract： We explore cross-domain offline reinforcement learning (RL) where offline datasets from another domain can be accessed to facilitate policy learning. However, the underlying environments of the two datasets may have dynamics mismatches, incurring inferior performance when simply merging the data of two domains. Existing methods mitigate this issue by training domain classifiers, using contrastive learning methods, etc. Nevertheless, they still rely on a large amount of target domain data to function well. Instead, we address this problem by establishing a concrete performance bound of a policy given datasets from two domains. Motivated by the theoretical insights, we propose to align transitions in the two datasets using optimal transport and selectively share source domain samples, without training any neural networks. This enables reliable data filtering even given a few target domain data. Additionally, we introduce a dataset regularization term that ensures the learned policy remains within the scope of the target domain dataset, preventing it from being biased towards the source domain data. Consequently, we propose the Optimal Transport Data Filtering (dubbed OTDF) method and examine its effectiveness by conducting extensive experiments across various dynamics shift conditions (e.g.,

gravity shift), given limited target domain data. It turns out that OTDF exhibits superior performance on many tasks and dataset qualities, often surpassing prior strong baselines by a large margin.

163. Revisiting Source-Free Domain Adaptation: a New Perspective via Uncertainty Control

链接：https://iclr.cc/virtual/2025/poster/28386 abstract： Source-Free Domain Adaptation (SFDA) seeks to adapt a pre-trained source model to the target domain using only unlabeled target data, without access to the original source data. While current state-of-the-art (SOTA) methods rely on leveraging weak supervision from the source model to extract reliable information for self-supervised adaptation, they often overlook the uncertainty that arises during the transfer process. In this paper, we conduct a systematic and theoretical analysis of the uncertainty inherent in existing SFDA methods and demonstrate its impact on transfer performance through the lens of Distributionally Robust Optimization (DRO). Building upon the theoretical results, we propose a novel instance-dependent uncertainty control algorithm for SFDA. Our method is designed to quantify and exploit the uncertainty during the adaptation process, significantly improving the model performance. Extensive experiments on benchmark datasets and empirical analyses confirm the validity of our theoretical findings and the effectiveness of the proposed method. This work offers new insights into understanding and advancing SFDA performance.

164. ESE: Espresso Sentence Embeddings

链接：https://iclr.cc/virtual/2025/poster/28282 abstract： High-quality sentence embeddings are fundamental in many natural language processing (NLP) tasks, such as semantic textual similarity (STS) and retrieval-augmented generation (RAG). However, most existing methods leverage fixed-length sentence embeddings from full-layer language models, which lack the scalability to accommodate the diverse available resources across various applications. Viewing this gap, we propose a novel sentence embedding model Espresso Sentence Embeddings (ESE) with two learning processes. First, the learn-to-express process encodes more salient representations to shallow layers. Second, the learn-to-compress process compacts essential features into the initial dimensions using Principal Component Analysis (PCA). This way, ESE can scale model depth via the former process and embedding size via the latter. Extensive experiments on STS and RAG suggest that ESE can effectively produce high-quality sentence embeddings with less model depth and embedding size, enhancing inference efficiency. The code is available at https://github.com/SeanLee97/AnglE/blob/main/README_ESE.md.

165. Simple is Effective: The Roles of Graphs and Large Language Models in Knowledge-Graph-Based Retrieval-Augmented Generation

链接：https://iclr.cc/virtual/2025/poster/30084 abstract： Large Language Models (LLMs) demonstrate strong reasoning abilities but face limitations such as hallucinations and outdated knowledge. Knowledge Graph (KG)-based Retrieval-Augmented Generation (RAG) addresses these issues by grounding LLM outputs in structured external knowledge from KGs. However, current KG-based RAG frameworks still struggle to optimize the trade-off between retrieval effectiveness and efficiency in identifying a suitable amount of relevant graph information for the LLM to digest. We introduce SubgraphRAG, extending the KG-based RAG framework that retrieves subgraphs and leverages LLMs for reasoning and answer prediction. Our approach innovatively integrates a lightweight multilayer perceptron (MLP) with a parallel triple-scoring mechanism for efficient and flexible subgraph retrieval while encoding directional structural distances to enhance retrieval effectiveness. The size of retrieved subgraphs can be flexibly adjusted to match the query's needs and the downstream LLM's capabilities. This design strikes a balance between model complexity and reasoning power, enabling scalable and generalizable retrieval processes. Notably, based on our retrieved subgraphs, smaller LLMs like Llama3.1-8B-Instruct deliver competitive results with explainable reasoning, while larger models like GPT-4o achieve state-of-the-art accuracy compared with previous baselines—all without fine-tuning. Extensive evaluations on the WebQSP and CWQ benchmarks highlight SubgraphRAG's strengths in efficiency, accuracy, and reliability by reducing hallucinations and improving response grounding.

166. Beyond Autoregression: Fast LLMs via Self-Distillation Through Time

链接：https://iclr.cc/virtual/2025/poster/27972 abstract： Autoregressive (AR) Large Language Models (LLMs) have demonstrated significant success across numerous tasks. However, the AR modeling paradigm presents certain limitations; for instance, contemporary autoregressive LLMs are trained to generate one token at a time, which can result in noticeable latency. Recent advances have indicated that search and repeated sampling can enhance performance in various applications, such as theorem proving, code generation, and alignment, by utilizing greater computational resources during inference. In this study, we demonstrate that diffusion language models are capable of generating at least 32 tokens simultaneously, while exceeding the performance of AR models in text quality and on the LAMBADA natural language understanding benchmark. This outcome is achieved through a novel distillation method for discrete diffusion models, which reduces the number of inference steps by a factor of 32-64. Practically, at the 1.3B parameters scale, diffusion models, even without caching, can generate tokens at a rate that is up to 8 times faster than AR models employing KV-caching, and we anticipate further improvements with the inclusion of caching. Moreover, we demonstrate the efficacy of our approach for diffusion language models with up to 860M parameters.

167. Streaming Video Understanding and Multi-round Interaction with Memory-enhanced Knowledge

链接：https://iclr.cc/virtual/2025/poster/30097 abstract： Recent advances in Large Language Models (LLMs) have enabled the development of Video-LLMs, advancing multimodal learning by bridging video data with language tasks. However, current video understanding models struggle with processing long video sequences, supporting multi-turn dialogues, and adapting to real-world dynamic scenarios. To address these issues, we propose StreamChat, a training-free framework for streaming video reasoning and conversational interaction. StreamChat leverages a novel hierarchical memory system to efficiently process and compress video features over extended sequences, enabling real-time, multi-turn dialogue. Our framework incorporates a parallel system scheduling strategy that enhances processing speed and reduces latency, ensuring robust performance in real-world applications. Furthermore, we introduce StreamBench, a versatile benchmark that evaluates streaming video understanding across diverse media types and interactive scenarios, including multi-turn interactions and complex reasoning tasks. Extensive evaluations on StreamBench and other public benchmarks demonstrate that StreamChat significantly outperforms existingstate-of-the-art models in terms of accuracy and response times, confirming its effectiveness for streaming video understanding. Code is available at StreamChat.

168. 3DIS: Depth-Driven Decoupled Image Synthesis for Universal Multi-Instance Generation

链接：https://iclr.cc/virtual/2025/poster/29926 abstract： The increasing demand for controllable outputs in text-to-image generation has spurred advancements in multi-instance generation (MIG), allowing users to define both instance layouts and attributes. However, unlike image-conditional generation methods such as ControlNet, MIG techniques have not been widely adopted in state-of-the-art models like SD2 and SDXL, primarily due to the challenge of building robust renderers that simultaneously handle instance positioning and attribute rendering. In this paper, we introduce Depth-Driven Decoupled Image Synthesis (3DIS), a novel framework that decouples the MIG process into two stages: (i) generating a coarse scene depth map for accurate instance positioning and scene composition, and (ii) rendering fine-grained attributes using pre-trained ControlNet on any foundational model, without additional training. Our 3DIS framework integrates a custom adapter into LDM3D for precise depth-based layouts and employs a finetuning-free method for enhanced instance-level attribute rendering. Extensive experiments on COCO-Position and COCO-MIG benchmarks demonstrate that 3DIS significantly outperforms existing methods in both layout precision and attribute rendering. Notably, 3DIS offers seamless compatibility with diverse foundational models, providing a robust, adaptable solution for advanced multi-instance generation. The code is available at: https://github.com/limuloo/3DIS.

169. ARLON: Boosting Diffusion Transformers with Autoregressive Models for Long Video Generation

链接：https://iclr.cc/virtual/2025/poster/30742 abstract： Text-to-video (T2V) models have recently undergone rapid and substantial advancements. Nevertheless, due to limitations in data and computational resources, achieving efficient generation of long videos with rich motion dynamics remains a significant challenge. To generate high-quality, dynamic, and temporally consistent long videos, this paper presents ARLON, a novel framework that boosts diffusion Transformers with autoregressive (\textbf{AR}) models for long (\textbf{LON}) video generation, by integrating the coarse spatial and long-range temporal information provided by the AR model to guide the DiT model effectively.Specifically, ARLON incorporates several key innovations: 1) A latent Vector Quantized Variational Autoencoder (VQ-VAE) compresses the input latent space of the DiT model into compact and highly quantized visual tokens, bridging the AR and DiT models and balancing the learning complexity and information density;2) An adaptive norm-based semantic injection module integrates the coarse discrete visual units from the AR model into the DiT model, ensuring effective guidance during video generation; 3) To enhance the tolerance capability of noise introduced from the AR inference, the DiT model is trained with coarser visual latent tokens incorporated with an uncertainty sampling module. Experimental results demonstrate that ARLON significantly outperforms the baseline OpenSora-V1.2 on eight out of eleven metrics selected from VBench, with notable improvements in dynamic degree and aesthetic quality, while delivering competitive results on the remaining three and simultaneously accelerating the generation process. In addition, ARLON achieves state-of-the-art performance in long video generation, outperforming other open-source models in this domain. Detailed analyses of the improvements in inference efficiency are presented, alongside a practical application that demonstrates the generation of long videos using progressive text prompts. Project page: \url{http://aka.ms/arlon}.

170. Catastrophic Failure of LLM Unlearning via Quantization

链接：https://iclr.cc/virtual/2025/poster/28528 abstract： Large language models (LLMs) have shown remarkable proficiency in generating text, benefiting from extensive training on vast textual corpora. However, LLMs may also acquire unwanted behaviors from the diverse and sensitive nature of their training data, which can include copyrighted and private content. Machine unlearning has been introduced as a viable solution to remove the influence of such problematic content without the need for costly and time-consuming retraining. This process aims to erase specific knowledge from LLMs while preserving as much model utility as possible. Despite the effectiveness of current unlearning methods, little attention has been given to whether existing unlearning methods for LLMs truly achieve forgetting or merely hide the knowledge, which current unlearning benchmarks fail to detect. This paper reveals that applying quantization to models that have undergone unlearning can restore the "forgotten" information. We conduct comprehensive experiments using various quantization techniques across multiple precision levels to thoroughly evaluate this phenomenon. We find that for unlearning methods with utility constraints, the unlearned model retains an average of 21\% of the intended forgotten knowledge in full precision, which significantly increases to 83\% after 4-bit quantization. Based on our empirical findings, we provide a theoretical explanation for the observed phenomenon and propose a quantization-robust unlearning strategy aimed at mitigating this intricate issue. Our results highlight

a fundamental tension between preserving the utility of the unlearned model and preventing knowledge recovery through quantization, emphasizing the challenge of balancing these two objectives. Altogether, our study underscores a major failure in existing unlearning methods for LLMs, strongly advocating for more comprehensive and robust strategies to ensure authentic unlearning without compromising model utility. Our code is available at: https://github.com/zzwjames/FailureLLMUnlearning.

171. Gaussian Head & Shoulders: High Fidelity Neural Upper Body Avatars with Anchor Gaussian Guided Texture Warping

链接：https://iclr.cc/virtual/2025/poster/30204 abstract： The ability to reconstruct realistic and controllable upper body avatars from casual monocular videos is critical for various applications in communication and entertainment. By equipping the most recent 3D Gaussian Splatting representation with head 3D morphable models (3DMM), existing methods manage to create head avatars with high fidelity. However, most existing methods only reconstruct a head without the body, substantially limiting their application scenarios. We found that naively applying Gaussians to model the clothed chest and shoulders tends to result in blurry reconstruction and noisy floaters under novel poses. This is because of the fundamental limitation of Gaussians and point clouds -- each Gaussian or point can only have a single directional radiance without spatial variance, therefore an unnecessarily large number of them is required to represent complicated spatially varying texture, even for simple geometry. In contrast, we propose to model the body part with a neural texture that consists of coarse and pose-dependent fine colors. To properly render the body texture for each view and pose without accurate geometry nor UV mapping, we optimize another sparse set of Gaussians as anchors that constrain the neural warping field that maps image plane coordinates to the texture space. We demonstrate that Gaussian Head & Shoulders can fit the high-frequency details on the clothed upper body with high fidelity and potentially improve the accuracy and fidelity of the head region. We evaluate our method with casual phone-captured and internet videos and show our method archives superior reconstruction quality and robustness in both self and cross reenactment tasks. To fully utilize the efficient rendering speed of Gaussian splatting, we additionally propose an accelerated inference method of our trained model without Multi-Layer Perceptron (MLP) queries and reach a stable rendering speed of around 130 FPS for any subjects.

172. Robustness Inspired Graph Backdoor Defense

链接：https://iclr.cc/virtual/2025/poster/28019 abstract： Graph Neural Networks (GNNs) have achieved promising results in tasks such as node classification and graph classification. However, recent studies reveal that GNNs are vulnerable to backdoor attacks, posing a significant threat to their real-world adoption. Despite initial efforts to defend against specific graph backdoor attacks, there is no work on defending against various types of backdoor attacks where generated triggers have different properties. Hence, we first empirically verify that prediction variance under edge dropping is a crucial indicator for identifying poisoned nodes. With this observation, we propose using random edge dropping to detect backdoors and theoretically show that it can efficiently distinguish poisoned nodes from clean ones. Furthermore, we introduce a novel robust training strategy to efficiently counteract the impact of the triggers. Extensive experiments on real-world datasets show that our framework can effectively identify poisoned nodes, significantly degrade the attack success rate, and maintain clean accuracy when defending against various types of graph backdoor attacks with different properties. Our code is available at: https://github.com/zzwjames/RIGBD.

173. Specialized Foundation Models Struggle to Beat Supervised Baselines

链接：https://iclr.cc/virtual/2025/poster/30102 abstract： Following its success for vision and text, the "foundation model" (FM) paradigm—pretraining large models on massive data, then fine-tuning on target tasks—has rapidly expanded to domains in the sciences, engineering, healthcare, and beyond. Has this achieved what the original FMs accomplished, i.e. the supplanting of traditional supervised learning in their domains? To answer we look at three modalities—genomics, satellite imaging, and time series—with multiple recent FMs and compare them to a standard supervised learning workflow: model development, hyperparameter tuning, and training, all using only data from the target task. Across these three specialized domains, we find that it is consistently possible to train simple supervised models—no more complicated than a lightly modified wide ResNet or UNet—that match or even outperform the latest foundation models. Our work demonstrates that the benefits of large-scale pretraining have yet to be realized in many specialized areas, reinforces the need to compare new FMs to strong, well-tuned baselines, and introduces two new, easy-to-use, open-source, and automated workflows for doing so.

174. R2Det: Exploring Relaxed Rotation Equivariance in 2D Object Detection

链接：https://iclr.cc/virtual/2025/poster/30402 abstract： Group Equivariant Convolution (GConv) empowers models to explore underlying symmetry in data, improving performance. However, real-world scenarios often deviate from ideal symmetric systems caused by physical permutation, characterized by non-trivial actions of a symmetry group, resulting in asymmetries that affect the outputs, a phenomenon known as Symmetry Breaking. Traditional GConv-based methods are constrained by rigid operational rules within group space, assuming data remains strictly symmetry after limited group transformations. This limitation makes it difficult to adapt to Symmetry-Breaking and non-rigid transformations. Motivated by this, we mainly focus on a common scenario: Rotational Symmetry-Breaking. By relaxing strict group transformations within Strict Rotation-Equivariant group $\mathbf{C}_n$, we redefine a Relaxed Rotation-Equivariant group $\mathbf{R}_n$ and introduce a novel Relaxed Rotation-Equivariant GConv (R2GConv) with only a minimal increase of $4n$ parameters compared to GConv. Based on R2GConv, we propose a Relaxed Rotation-Equivariant Network (R2Net) as the backbone and develop a Relaxed Rotation-Equivariant Object Detector (R2Det)

for 2D object detection. Experimental results demonstrate the effectiveness of the proposed R2GConv in natural image classification, and R2Det achieves excellent performance in 2D object detection with improved generalization capabilities and robustness. The code is available in \texttt{https://github.com/wuer5/r2det}.

175. Do You Keep an Eye on What I Ask? Mitigating Multimodal Hallucination via Attention-Guided Ensemble Decoding

链接：https://iclr.cc/virtual/2025/poster/27655 abstract： Recent advancements in Large Vision-Language Models (LVLMs) have significantly expanded their utility in tasks like image captioning and visual question answering. However, they still struggle with object hallucination, where models generate descriptions that inaccurately reflect the visual content by including nonexistent objects or misrepresenting existing ones. While previous methods, such as data augmentation and training-free approaches, strive to tackle this issue, they still encounter scalability challenges and often depend on additional external modules. In this work, we propose Ensemble Decoding (ED), a novel strategy that splits the input image into sub-images and combines logit distributions by assigning weights through the attention map. Furthermore, we introduce ED adaptive plausibility constraint to calibrate logit distribution and FastED, a variant designed for speed-critical applications. Extensive experiments across hallucination benchmarks demonstrate that our proposed method achieves state-of-the-art performance, validating the effectiveness of our approach.

176. LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations

链接：https://iclr.cc/virtual/2025/poster/30060 abstract： Large language models (LLMs) often produce errors, including factual inaccuracies, biases, and reasoning failures, collectively referred to as "hallucinations". Recent studies have demonstrated that LLMs' internal states encode information regarding the truthfulness of their outputs, and that this information can be utilized to detect errors. In this work, we show that the internal representations of LLMs encode much more information about truthfulness than previously recognized. We first discover that the truthfulness information is concentrated in specific tokens, and leveraging this property significantly enhances error detection performance. Yet, we show that such error detectors fail to generalize across datasets, implying that---contrary to prior claims---truthfulness encoding is not universal but rather multifaceted. Next, we show that internal representations can also be used for predicting the types of errors the model is likely to make, facilitating the development of tailored mitigation strategies. Lastly, we reveal a discrepancy between LLMs' internal encoding and external behavior: they may encode the correct answer, yet consistently generate an incorrect one. Taken together, these insights deepen our understanding of LLM errors from the model's internal perspective, which can guide future research on enhancing error analysis and mitigation.

177. NL-Eye: Abductive NLI For Images

链接：https://iclr.cc/virtual/2025/poster/31098 abstract： Will a Visual Language Model (VLM)-based bot warn us about slipping if it detects a wet floor? Recent VLMs have demonstrated impressive capabilities, yet their ability to infer outcomes and causes remains underexplored. To address this, we introduce NL-Eye, a benchmark designed to assess VLMs' visual abductive reasoning skills. NL-Eye adapts the abductive Natural Language Inference (NLI) task to the visual domain, requiring models to evaluate the plausibility of hypothesis images based on a premise image and explain their decisions. NL-Eye consists of 350 carefully curated triplet examples (1,050 images) spanning diverse reasoning categories: physical, functional, logical, emotional, cultural, and social. The data curation process involved two steps—writing textual descriptions and generating images using text-to-image models, both requiring substantial human involvement to ensure high-quality and challenging scenes. Our experiments show that VLMs struggle significantly on NL-Eye, often performing at random baseline levels, while humans excel in both plausibility prediction and explanation quality. This demonstrates a deficiency in the abductive reasoning capabilities of modern VLMs. NL-Eye represents a crucial step toward developing VLMs capable of robust multimodal reasoning for real-world applications, including accident-prevention bots and generated video verification.

178. Lost in Prediction: Why Social Media Narratives Don't Help Macroeconomic Forecasting?

链接：https://iclr.cc/virtual/2025/poster/31323 abstract： Can we predict the macroeconomy by analyzing the narratives people share on social media? We dove deep into the world of Narrative Economics, using NLP models to analyze millions of viral tweets and see if they could help us predict the fluctuations of macroeconomic indicators. □ Spoiler alert: it's not that easy! Join us as we explore the interesting relationship between narratives, social media, and macroeconomy, and uncover the challenges of turning narratives into treasure.

179. Exploring the Camera Bias of Person Re-identification

链接：https://iclr.cc/virtual/2025/poster/29586 abstract： We empirically investigate the camera bias of person re-identification (ReID) models. Previously, camera-aware methods have been proposed to address this issue, but they are largely confined to training domains of the models. We measure the camera bias of ReID models on unseen domains and reveal that camera bias becomes more pronounced under data distribution shifts. As a debiasing method for unseen domain data, we revisit feature

normalization on embedding vectors. While the normalization has been used as a straightforward solution, its underlying causes and broader applicability remain unexplored. We analyze why this simple method is effective at reducing bias and show that it can be applied to detailed bias factors such as low-level image properties and body angle. Furthermore, we validate its generalizability across various models and benchmarks, highlighting its potential as a simple yet effective test-time postprocessing method for ReID. In addition, we explore the inherent risk of camera bias in unsupervised learning of ReID models. The unsupervised models remain highly biased towards camera labels even for seen domain data, indicating substantial room for improvement. Based on observations of the negative impact of camera-biased pseudo labels on training, we suggest simple training strategies to mitigate the bias. By applying these strategies to existing unsupervised learning algorithms, we show that significant performance improvements can be achieved with minor modifications.

180. Improving Probabilistic Diffusion Models With Optimal Diagonal Covariance Matching

链接：https://iclr.cc/virtual/2025/poster/28877 abstract： The probabilistic diffusion model has become highly effective across various domains. Typically, sampling from a diffusion model involves using a denoising distribution characterized by a Gaussian with a learned mean and either fixed or learned covariances. In this paper, we leverage the recently proposed covariance moment matching technique and introduce a novel method for learning the diagonal covariances. Unlike traditional data-driven covariance approximation approaches, our method involves directly regressing the optimal analytic covariance using a new, unbiased objective named Optimal Covariance Matching (OCM). This approach can significantly reduce the approximation error in covariance prediction. We demonstrate how our method can substantially enhance the sampling efficiency, recall rate and likelihood of both diffusion models and latent diffusion models.

181. Multimodal Large Language Models for Inverse Molecular Design with Retrosynthetic Planning

链接：https://iclr.cc/virtual/2025/poster/28198 abstract： While large language models (LLMs) have integrated images, adapting them to graphs remains challenging, limiting their applications in materials and drug design. This difficulty stems from the need for coherent autoregressive generation across texts and graphs. To address this, we introduce Llamole, the first multimodal LLM capable of interleaved text and graph generation, enabling molecular inverse design with retrosynthetic planning. Llamole integrates a base LLM with the Graph Diffusion Transformer and Graph Neural Networks for multi-conditional molecular generation and reaction inference within texts, while the LLM, with enhanced molecular understanding, flexibly controls activation among the different graph modules. Additionally, Llamole integrates A* search with LLM-based cost functions for efficient retrosynthetic planning. We create benchmarking datasets and conduct extensive experiments to evaluate Llamole against in-context learning and supervised fine-tuning. Llamole significantly outperforms 14 adapted LLMs across 12 metrics for controllable molecular design and retrosynthetic planning. Code and model at https://github.com/liugangcode/Llamole.

182. Towards Foundation Models for Mixed Integer Linear Programming

链接：https://iclr.cc/virtual/2025/poster/30856 abstract： Mixed Integer Linear Programming (MILP) is essential for modeling complex decision-making problems but faces challenges in computational tractability and interpretability. Current deep learning approaches for MILP focus on specific problem classes and do not generalize to unseen classes. To address this shortcoming, we take a foundation model training approach, where we train a single deep learning model on a diverse set of MILP problems to generalize across problem classes. As existing datasets for MILP lack diversity and volume, we introduce MILP-Evolve, a novel LLM-based evolutionary framework that is capable of generating a large set of diverse MILP classes with an unlimited amount of instances. We study our methodology on three key learning tasks that capture diverse aspects of MILP: (1) integrality gap prediction, (2) learning to branch, and (3) a new task of aligning MILP instances with natural language descriptions. Our empirical results show that models trained on the data generated by MILP-Evolve achieve significant improvements on unseen problems, including MIPLIB benchmarks. Our work highlights the potential of moving towards a foundation model approach for MILP that can generalize to a broad range of MILP problem classes. Our code and data are publicly available at https://github.com/microsoft/OptiGuide.

183. Your Absorbing Discrete Diffusion Secretly Models the Conditional Distributions of Clean Data

链接：https://iclr.cc/virtual/2025/poster/28131 abstract： Discrete diffusion models with absorbing processes have shown promise in language modeling. The key quantities to be estimated are the ratios between the marginal probabilities of two transitive states at all timesteps, called the concrete score. In this paper, we reveal that the concrete score in absorbing diffusion can be expressed as conditional probabilities of clean data, multiplied by a time-dependent scalar in an analytic form. Motivated by this finding, we propose reparameterized absorbing discrete diffusion (RADD), a dedicated diffusion model without time-condition that characterizes the time-independent conditional probabilities. Besides its simplicity, RADD can reduce the number of function evaluations (NFEs) by caching the output of the time-independent network when the noisy sample remains unchanged in a sampling interval, which enables sampling acceleration. Built upon the new perspective of conditional distributions, we further unify absorbing discrete diffusion and any-order autoregressive models (AO-ARMs), showing that the upper bound on the negative log-likelihood for the diffusion model can be interpreted as an expected negative log-likelihood for AO-ARMs. Further,

our RADD models achieve SOTA performance among diffusion models on 5 zero-shot language modeling benchmarks (measured by perplexity) at the GPT-2 scale. Our code is available at \url{https://github.com/ML-GSAI/RADD}.

184. Second Order Bounds for Contextual Bandits with Function Approximation

链接：https://iclr.cc/virtual/2025/poster/28780 abstract： Many works have developed no-regret algorithms for contextual bandits with function approximation, where the mean rewards over context-action pairs belong to a function class $\mathcal{F}$. Although there are many approaches to this problem, algorithms based on the principle of optimism, such as optimistic least squares have gained in importance. It can be shown the regret of this algorithm scales as $\widetilde{\mathcal{O}}\left(\sqrt{d_{\mathrm{eluder}}(\mathcal{F}) \log(\mathcal{F}) T }\right)$ where $d_{\mathrm{eluder}}(\mathcal{F})$ is a statistical measure of the complexity of the function class $\mathcal{F}$ known as eluder dimension. Unfortunately, even if the variance of the measurement noise of the rewards at time t equals σ_t^2 and these are close to zero, the optimistic least squares algorithm’s regret scales with $\sqrt{T}$. In this work we are the first to develop algorithms that satisfy regret bounds for contextual bandits with function approximation of the form $\widetilde{\mathcal{O}}\left(\sigma \sqrt{\log(\mathcal{F})d_{\mathrm{eluder}}(\mathcal{F}) T } + d_{\mathrm{eluder}}(\mathcal{F}) \cdot \log(|\mathcal{F}|)\right) $ when the variances are unknown and satisfy $\sigma_t^2 = \sigma$ for all t and $\widetilde{\mathcal{O}}\left(d_{\mathrm{eluder}}(\mathcal{F})\sqrt{\log(\mathcal{F})\sum_{t=1}^T \sigma_t^2 } + d_{\mathrm{eluder}}(\mathcal{F}) \cdot \log(|\mathcal{F}|)\right) $ when the variances change every time-step. These bounds generalize existing techniques for deriving second order bounds in contextual linear problems.

185. A Spark of Vision-Language Intelligence: 2-Dimensional Autoregressive Transformer for Efficient Finegrained Image Generation

链接：https://iclr.cc/virtual/2025/poster/27817 abstract： This work tackles the information loss bottleneck of vector-quantization (VQ) autoregressive image generation by introducing a novel model architecture called the 2-Dimensional Autoregression (DnD) Transformer. The DnD-Transformer predicts more codes for an image by introducing a new direction, model depth, along with the sequence length. Compared to 1D autoregression and previous work using similar 2D image decomposition such as RQ-Transformer, the DnD-Transformer is an end-to-end model that can generate higher quality images with the same backbone model size and sequence length, opening a new optimization perspective for autoregressive image generation. Furthermore, our experiments reveal that the DnD-Transformer's potential extends beyond generating natural images. It can even generate images with rich text and graphical elements in a self-supervised manner, demonstrating an understanding of these combined modalities. This has not been previously demonstrated for popular vision generative models such as diffusion models, showing a spark of vision-language intelligence when trained solely on images. Code, datasets and models are open at https://github.com/chenllliang/DnD-Transformer.

186. TetSphere Splatting: Representing High-Quality Geometry with Lagrangian Volumetric Meshes

链接：https://iclr.cc/virtual/2025/poster/30753 abstract： We introduce TetSphere Splatting, a Lagrangian geometry representation designed for high-quality 3D shape modeling. TetSphere splatting leverages an underused yet powerful geometric primitive -- volumetric tetrahedral meshes. It represents 3D shapes by deforming a collection of tetrahedral spheres, with geometric regularizations and constraints that effectively resolve common mesh issues such as irregular triangles, non-manifoldness, and floating artifacts. Experimental results on multi-view and single-view reconstruction highlight TetSphere splatting's superior mesh quality while maintaining competitive reconstruction accuracy compared to state-of-the-art methods. Additionally, TetSphere splatting demonstrates versatility by seamlessly integrating into generative modeling tasks, such as image-to-3D and text-to-3D generation.

187. Wavelet-based Positional Representation for Long Context

链接：https://iclr.cc/virtual/2025/poster/29808 abstract： In the realm of large-scale language models, a significant challenge arises when extrapolating sequences beyond the maximum allowable length. This is because the model's position embedding mechanisms are limited to positions encountered during training, thus preventing effective representation of positions in longer sequences.We analyzed conventional position encoding methods for long contexts and found the following characteristics.(1) When the representation dimension is regarded as the time axis, Rotary Position Embedding (RoPE) can be interpreted as a restricted wavelet transform using Haar-like wavelets. However, because it uses only a fixed scale parameter, it does not fully exploit the advantages of wavelet transforms, which capture the fine movements of non-stationary signals using multiple scales (window sizes). This limitation could explain why RoPE performs poorly in extrapolation.(2)Previous research as well as our own analysis indicates that Attention with Linear Biases (ALiBi) functions similarly to windowed attention, using windows of varying sizes.However, it has limitations in capturing deep dependencies because it restricts the receptive field of the model.From these insights, we propose a new position representation method that captures multiple scales (i.e., window sizes) by leveraging wavelet transforms without limiting the model's attention field.Experimental results show that this new method improves the performance of the model in both short and long contexts. In particular, our method allows extrapolation of position information without limiting the model's attention field.

188. Inverse Scaling: When Bigger Isn't Better

链接：https://iclr.cc/virtual/2025/poster/31511 abstract： Work on scaling laws has found that large language models (LMs) show predictable improvements to overall loss with increased scale (model size, training data, and compute). Here, we present evidence for the claim that LMs may show inverse scaling, or worse task performance with increased scale, e.g., due to flaws in the training objective and data. We present empirical evidence of inverse scaling on 11 datasets collected by running a public contest, the Inverse Scaling Prize, with a substantial prize pool. Through analysis of the datasets, along with other examples found in the literature, we identify four potential causes of inverse scaling: (i) preference to repeat memorized sequences over following in-context instructions, (ii) imitation of undesirable patterns in the training data, (iii) tasks containing an easy distractor task which LMs could focus on, rather than the harder real task, and (iv) correct but misleading few-shot demonstrations of the task. We release the winning datasets at inversescaling.com/data to allow for further investigation of inverse scaling. Our tasks have helped drive the discovery of U-shaped and inverted-U scaling trends, where an initial trend reverses, suggesting that scaling trends are less reliable at predicting the behavior of larger-scale models than previously understood. Overall, our results suggest that there are tasks for which increased model scale alone may not lead to progress, and that more careful thought needs to go into the data and objectives for training language models.

189. Procedural Synthesis of Synthesizable Molecules

链接：https://iclr.cc/virtual/2025/poster/29833 abstract： Designing synthetically accessible molecules and recommending analogs to unsynthesizable molecules are important problems for accelerating molecular discovery. We reconceptualize both problems using ideas from program synthesis. Drawing inspiration from syntax-guided synthesis approaches, we decouple the syntactic skeleton from the semantics of a synthetic tree to create a bilevel framework for reasoning about the combinatorial space of synthesis pathways. Given a molecule we aim to generate analogs for, we iteratively refine its skeletal characteristics via Markov Chain Monte Carlo simulations over the space of syntactic skeletons. Given a black-box oracle to optimize, we formulate a joint design space over syntactic templates and molecular descriptors and introduce evolutionary algorithms that optimize both syntactic and semantic dimensions synergistically. Our key insight is that once the syntactic skeleton is set, we can amortize over the search complexity of deriving the program's semantics by training policies to fully utilize the fixed horizon Markov Decision Process imposed by the syntactic template. We demonstrate performance advantages of our bilevel framework for synthesizable analog generation and synthesizable molecule design. Notably, our approach offers the user explicit control over the resources required to perform synthesis and biases the design space towards simpler solutions, making it particularly promising for autonomous synthesis platforms. Supporting code is at https://github.com/shiningsunnyday/SynthesisNet.

190. Dynamic-SUPERB Phase-2: A Collaboratively Expanding Benchmark for Measuring the Capabilities of Spoken Language Models with 180 Tasks

链接：https://iclr.cc/virtual/2025/poster/32056 abstract： Multimodal foundation models, such as Gemini and ChatGPT, have revolutionized human-machine interactions by seamlessly integrating various forms of data. Developing a universal spoken language model that comprehends a wide range of natural language instructions is critical for bridging communication gaps and facilitating more intuitive interactions. However, the absence of a comprehensive evaluation benchmark poses a significant challenge. We present Dynamic-SUPERB Phase-2, an open and evolving benchmark for the comprehensive evaluation of instruction-based universal speech models. Building upon the first generation, this second version incorporates 125 new tasks contributed collaboratively by the global research community, expanding the benchmark to a total of 180 tasks, making it the largest benchmark for speech and audio evaluation. While the first generation of Dynamic-SUPERB was limited to classification tasks, Dynamic-SUPERB Phase-2 broadens its evaluation capabilities by introducing a wide array of novel and diverse tasks, including regression and sequence generation, across speech, music, and environmental audio. Evaluation results show that no model performed well universally. SALMONN-13B excelled in English ASR and Qwen2-Audio-7B-Instruct showed high accuracy in emotion recognition, but current models still require further innovations to handle a broader range of tasks. We open-source all task data and the evaluation pipeline at https://github.com/dynamic-superb/dynamic-superb.

191. LOIRE: LifelOng learning on Incremental data via pre-trained language model gRowth Efficiently

链接：https://iclr.cc/virtual/2025/poster/30366 abstract： Large-scale pre-trained language models (PLMs) require significant computational resources to train from scratch on large volumes of data. But in the real world, emerging data from diverse sources may not be initially available for pre-training. Recent studies on lifelong learning have tried to solve this problem by exploring the use of model growth techniques to effectively incorporate new knowledge without the need for complete re-training. However, model growth approaches utilized have issues with growth operators that do not ensure strict function preservation or growth schedules that only include a few growth dimensions, reducing lifelong learning's effect. Furthermore, existing approaches often assume that emerging data has the same distribution as pre-training data, causing catastrophic forgetting of previously acquired knowledge. To address the aforementioned issues, we introduce LOIRE, a framework for lifelong learning that enables PLMs to effectively grow their capacity using incremental data. LOIRE employs growth operators for all feasible dimensions and a growth schedule to generate the optimal expansion sequence in the field of lifelong learning. Specifically, we present a novel plug-in layer growth operator with residual connections that skip the newly added layer during initial training while ensuring function preservation. We additionally propose an iterative distillation strategy for LOIRE that allows an intermediate

model in the growth stages to switch between being a student and a teacher, reducing catastrophic forgetting during growth. Experiments show that LOIRE can reduce computational expenses by an average of 29.22\% while retaining equivalent or better downstream performance.

192. Conditional Testing based on Localized Conformal p-values

链接：https://iclr.cc/virtual/2025/poster/30148 abstract： In this paper, we address conditional testing problems through the conformal inference framework. We define the localized conformal p-values by inverting prediction intervals and prove their theoretical properties. These defined p-values are then applied to several conditional testing problems to illustrate their practicality. Firstly, we propose a conditional outlier detection procedure to test for outliers in the conditional distribution with finite-sample false discovery rate (FDR) control. We also introduce a novel conditional label screening problem with the goal of screening multivariate response variables and propose a screening procedure to control the family-wise error rate (FWER). Finally, we consider the two-sample conditional distribution test and define a weighted U-statistic through the aggregation of localized p-values. Numerical simulations and real-data examples validate the superior performance of our proposed strategies.

193. Error-quantified Conformal Inference for Time Series

链接：https://iclr.cc/virtual/2025/poster/29659 abstract： Uncertainty quantification in time series prediction is challenging due to the temporal dependence and distribution shift on sequential data. Conformal prediction provides a pivotal and flexible instrument for assessing the uncertainty of machine learning models through prediction sets. Recently, a series of online conformal inference methods updated thresholds of prediction sets by performing online gradient descent on a sequence of quantile loss functions. A drawback of such methods is that they only use the information of revealed non-conformity scores via miscoverage indicators but ignore error quantification, namely the distance between the non-conformity score and the current threshold. To accurately leverage the dynamic of miscoverage error, we propose Error-quantified Conformal Inference (ECI) by smoothing the quantile loss function. ECI introduces a continuous and adaptive feedback scale with the miscoverage error, rather than simple binary feedback in existing methods. We establish a long-term coverage guarantee for ECI under arbitrary dependence and distribution shift. The extensive experimental results show that ECI can achieve valid miscoverage control and output tighter prediction sets than other baselines.

194. Occlusion-aware Non-Rigid Point Cloud Registration via Unsupervised Neural Deformation Correntropy

链接：https://iclr.cc/virtual/2025/poster/29032 abstract： Non-rigid alignment of point clouds is crucial for scene understanding, reconstruction, and various computer vision and robotics tasks. Recent advancements in implicit deformation networks for non-rigid registration have significantly reduced the reliance on large amounts of annotated training data. However, existing state-of-the-art methods still face challenges in handling occlusion scenarios. To address this issue, this paper introduces an innovative unsupervised method called Occlusion-Aware Registration (OAR) for non-rigidly aligning point clouds. The key innovation of our method lies in the utilization of the adaptive correntropy function as a localized similarity measure, enabling us to treat individual points distinctly. In contrast to previous approaches that solely minimize overall deviations between two shapes, we combine unsupervised implicit neural representations with the maximum correntropy criterion to optimize the deformation of unoccluded regions. This effectively avoids collapsed, tearing, and other physically implausible results. Moreover, we present a theoretical analysis and establish the relationship between the maximum correntropy criterion and the commonly used Chamfer distance, highlighting that the correntropy-induced metric can be served as a more universal measure for point cloud analysis. Additionally, we introducelocally linear reconstruction to ensure that regions lacking correspondences between shapes still undergo physically natural deformations. Our method achieves superior or competitive performance compared to existing approaches, particularly when dealing with occluded geometries. We also demonstrate the versatility of our method in challenging tasks such as large deformations, shape interpolation, and shape completion under occlusion disturbances.

195. VLMaterial: Procedural Material Generation with Large Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/27856 abstract： Procedural materials, represented as functional node graphs, are ubiquitous in computer graphics for photorealistic material appearance design. They allow users to perform intuitive and precise editing to achieve desired visual appearances. However, creating a procedural material given an input image requires professional knowledge and significant effort. In this work, we leverage the ability to convert procedural materials into standard Python programs and fine-tune a large pre-trained vision-language model (VLM) to generate such programs from input images. To enable effective fine-tuning, we also contribute an open-source procedural material dataset and propose to perform program-level augmentation by prompting another pre-trained large language model (LLM). Through extensive evaluation, we show that our method outperforms previous methods on both synthetic and real-world examples.

196. LoCA: Location-Aware Cosine Adaptation for Parameter-Efficient Fine-Tuning

链接：https://iclr.cc/virtual/2025/poster/31017 abstract： Low-rank adaptation (LoRA) has become a prevalent method for adapting pre-trained large language models to downstream tasks. However, the simple low-rank decomposition form may constrain the optimization flexibility. To address this limitation, we introduce Location-aware Cosine Adaptation (LoCA), a novel frequency-domain parameter-efficient fine-tuning method based on inverse Discrete Cosine Transform (iDCT) with selective locations of learnable components. We begin with a comprehensive theoretical comparison between frequency-domain and low-rank decompositions for fine-tuning pre-trained large models. Our analysis reveals that frequency-domain approximation with carefully selected frequency components can surpass the expressivity of traditional low-rank-based methods. Furthermore, we demonstrate that iDCT offers a more efficient implementation compared to inverse Discrete Fourier Transform (iDFT), allowing for better selection and tuning of frequency components while maintaining equivalent expressivity to the optimal iDFT-based adaptation. By employing finite-difference approximation to estimate gradients for discrete locations of learnable coefficients on the DCT spectrum, LoCA dynamically selects the most informative frequency components during training. Experiments on diverse language and vision fine-tuning tasks demonstrate that LoCA offers enhanced parameter efficiency while maintains computational feasibility comparable to low-rank-based methods.

197. Flow-based Variational Mutual Information: Fast and Flexible Approximations

链接：https://iclr.cc/virtual/2025/poster/28097 abstract： Mutual Information (MI) is a fundamental measure of dependence between random variables, but its practical application is limited because it is difficult to calculate in many circumstances. Variational methods offer one approach by introducing an approximate distribution to create various bounds on MI, which in turn is an easier optimization problem to solve. In practice, the variational distribution chosen is often a Gaussian, which is convenient but lacks flexibility in modeling complicated distributions. In this paper, we introduce new classes of variational estimators based on Normalizing Flows that extend the previous Gaussian-based variational estimators. Our new estimators maintain many of the same theoretical guarantees while simultaneously enhancing the expressivity of the variational distribution. We experimentally verify that our new methods are effective on large MI problems where discriminative-based estimators, such as MINE and InfoNCE, are fundamentally limited. Furthermore, we compare against a diverse set of benchmarking tests to show that the flow-based estimators often perform as well, if not better, than the discriminative-based counterparts. Finally, we demonstrate how these estimators can be effectively utilized in the Bayesian Optimal Experimental Design setting for online sequential decision making.

198. SMI-Editor: Edit-based SMILES Language Model with Fragment-level Supervision

链接：https://iclr.cc/virtual/2025/poster/29960 abstract： SMILES, a crucial textual representation of molecular structures, has garnered significant attention as a foundation for pre-trained language models (LMs). However, most existing pre-trained SMILES LMs focus solely on the single-token level supervision during pre-training, failing to fully leverage the substructural information of molecules. This limitation makes the pre-training task overly simplistic, preventing the models from capturing richer molecular semantic information. Moreover, during pre-training, these SMILES LMs only process corrupted SMILES inputs, never encountering any valid SMILES, which leads to a train-inference mismatch. To address these challenges, we propose SMI-Editor, a novel edit-based pre-trained SMILES LM. SMI-Editor disrupts substructures within a molecule at random and feeds the resulting SMILES back into the model, which then attempts to restore the original SMILES through an editing process. This approach not only introduces fragment-level training signals, but also enables the use of valid SMILES as inputs, allowing the model to learn how to reconstruct complete molecules from these incomplete structures. As a result, the model demonstrates improved scalability and an enhanced ability to capture fragment-level molecular information. Experimental results show that SMI-Editor achieves state-of-the-art performance across multiple downstream molecular tasks, and even outperforming several 3D molecular representation models.

199. Beyond the convexity assumption: Realistic tabular data generation under quantifier-free real linear constraints

链接：https://iclr.cc/virtual/2025/poster/28159 abstract： Synthetic tabular data generation has traditionally been a challenging problem due to the high complexity of the underlying distributions that characterise this type of data. Despite recent advances in deep generative models (DGMs), existing methods often fail to produce realistic datapoints that are well-aligned with available background knowledge.In this paper, we address this limitation by introducing Disjunctive Refinement Layer (DRL), a novel layer designedto enforce the alignment of generated data with the background knowledge specified in user-defined constraints.DRL is the first method able to automatically make deep learning models inherently compliant with constraints as expressive as quantifier-free linear formulas, which can define non-convex and even disconnected spaces. Our experimental analysis shows that DRL not only guarantees constraint satisfaction but also improves efficacy in downstream tasks. Notably, when applied to DGMs that frequently violate constraints, DRL eliminates violations entirely. Further, it improves performance metrics by up to 21.4\% in F1-score and 20.9\% in Area Under the ROC Curve, thus demonstrating its practical impact on data generation.

200. 3DitScene: Editing Any Scene via Language-guided Disentangled Gaussian Splatting

链接：https://iclr.cc/virtual/2025/poster/28702 abstract： Scene image editing is crucial for entertainment, photography, and advertising design. Existing methods solely focus on either 2D individual object or 3D global scene editing. This results in a lack of a unified approach to effectively control and manipulate scenes at the 3D level with different levels of granularity. In this work, we propose 3DitScene, a novel and unified scene editing framework leveraging language-guided disentangled Gaussian Splatting that enables seamless editing from 2D to 3D, allowing precise control over scene composition and individual objects. We first incorporate 3D Gaussians that are refined through generative priors and optimization techniques. Language features from CLIP then introduce semantics into 3D geometry for object disentanglement. With the disentangled Gaussians, 3DitScene allows for manipulation at both the global and individual levels, revolutionizing creative expression and empowering control over scenes and objects. Experimental results demonstrate the effectiveness and versatility of 3DitScene in scene image editing.