2801. Fully-inductive Node Classification on Arbitrary Graphs

链接：https://iclr.cc/virtual/2025/poster/31202 abstract： One fundamental challenge in graph machine learning is generalizing to new graphs. Many existing methods following the inductive setup can generalize to test graphs with new structures, but assuming the feature and label spaces remain the same as the training ones. This paper introduces a fully-inductive setup, where models should perform inference on arbitrary test graphs with new structures, feature and label spaces. We propose GraphAny as the first attempt at this challenging setup. GraphAny models inference on a new graph as an analytical solution to a LinearGNN, which can be naturally applied to graphs with any feature and label spaces. To further build a stronger model with learning capacity, we fuse multiple LinearGNN predictions with learned inductive attention scores. Specifically, the attention module is carefully parameterized as a function of the entropy-normalized distance features between pairs of LinearGNN predictions to ensure generalization to new graphs. Empirically, GraphAny trained on a single Wisconsin dataset with only 120 labeled nodes can generalize to 30 new graphs with an average accuracy of 67.26%, surpassing not only all inductive baselines, but also strong transductive methods trained separately on each of the 30 test graphs.

2802. On Large Language Model Continual Unlearning

链接：https://iclr.cc/virtual/2025/poster/30382 abstract： While large language models have demonstrated impressive performance across various domains and tasks, their security issues have become increasingly severe. Machine unlearning has emerged as a representative approach for model safety and security by removing the influence of undesired data on the target model. However, these methods do not sufficiently consider that unlearning requests in real-world scenarios are continuously emerging, especially in the context of LLMs, which may lead to accumulated model utility loss that eventually becomes unacceptable.Moreover, existing LLM unlearning methods often ignore previous data access limitations due to privacy concerns and copyright protection. Without previous data, the utility preservation during unlearning is much harder. To overcome these challenges, we propose the \OOO{} framework that includes an \underline{\textit{O}}rthogonal low-rank adapter (LoRA) for continually unlearning requested data and an \underline{\textit{O}}ut-\underline{\textit{O}}f-Distribution (OOD) detector to measure the similarity between input and unlearning data. The orthogonal LoRA achieves parameter disentanglement among continual unlearning requests. The OOD detector is trained with a novel contrastive entropy loss and utilizes a glocal-aware scoring mechanism. During inference, our \OOO{} framework can decide whether and to what extent to load the unlearning LoRA based on the OOD detector's predicted similarity between the input and the unlearned knowledge. Notably, \OOO{}'s effectiveness does not rely on any retained data. We conducted extensive experiments on \OOO{} and state-of-the-art LLM unlearning methods across three tasks and seven datasets. The results indicate that \OOO{} consistently achieves the best unlearning effectiveness and utility preservation, especially when facing continuous unlearning requests. The source codes can be found at \url{https://github.com/GCYZSL/O3-LLM-UNLEARNING}.

2803. MolSpectra: Pre-training 3D Molecular Representation with Multi-modal Energy Spectra

链接：https://iclr.cc/virtual/2025/poster/27793 abstract： Establishing the relationship between 3D structures and the energy states of molecular systems has proven to be a promising approach for learning 3D molecular representations. However, existing methods are limited to modeling the molecular energy states from classical mechanics. This limitation results in a significant oversight of quantum mechanical effects, such as quantized (discrete) energy level structures, which offer a more accurate estimation of molecular energy and can be experimentally measured through energy spectra. In this paper, we propose to utilize the energy spectra to enhance the pre-training of 3D molecular representations (MolSpectra), thereby infusing the knowledge of quantum mechanics into the molecular representations. Specifically, we propose SpecFormer, a multi-spectrum encoder for encoding molecular spectra via masked patch reconstruction. By further aligning outputs from the 3D encoder and spectrum encoder using a contrastive objective, we enhance the 3D encoder's understanding of molecules. Evaluations on public benchmarks reveal that our pre-trained representations surpass existing methods in predicting molecular properties and modeling dynamics.

2804. Jailbreak Antidote: Runtime Safety-Utility Balance via Sparse Representation Adjustment in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28150 abstract： As large language models (LLMs) become integral to various applications, ensuring both their safety and utility is paramount. Jailbreak attacks, which manipulate LLMs into generating harmful content, pose significant challenges to this balance. Existing defenses, such as prompt engineering and safety fine-tuning, often introduce computational overhead, increase inference latency, and lack runtime flexibility. Moreover, overly restrictive safety measures can degrade model utility by causing refusals of benign queries. In this paper, we introduce Jailbreak Antidote, a method that enables real-time adjustment of LLM safety preferences by manipulating a sparse subset of the model's internal states during inference. By shifting the model's hidden representations along a safety direction with varying strengths, we achieve flexible control over the safety-utility balance without additional token overhead or inference delays. Our analysis reveals that safety-related information in LLMs is sparsely distributed; adjusting approximately 5\% of the internal state is as effective as modifying the entire state. Extensive experiments on nine LLMs (ranging from 2 billion to 72 billion parameters), evaluated against ten jailbreak attack methods and compared with six defense strategies, validate the effectiveness and efficiency of our approach. By directly manipulating internal states during reasoning, Jailbreak Antidote offers a lightweight, scalable solution that enhances LLM safety while preserving utility, opening new possibilities for real-time safety mechanisms in widely-deployed AI

systems.

2805. LayerDAG: A Layerwise Autoregressive Diffusion Model for Directed Acyclic Graph Generation

链接：https://iclr.cc/virtual/2025/poster/31518 abstract： Directed acyclic graphs (DAGs) serve as crucial data representations in domains such as hardware synthesis and compiler/program optimization for computing systems. DAG generative models facilitate the creation of synthetic DAGs, which can be used for benchmarking computing systems while preserving intellectual property. However, generating realistic DAGs is challenging due to their inherent directional and logical dependencies. This paper introduces LayerDAG, an autoregressive diffusion model, to address these challenges. LayerDAG decouples the strong node dependencies into manageable units that can be processed sequentially. By interpreting the partial order of nodes as a sequence of bipartite graphs, LayerDAG leverages autoregressive generation to model directional dependencies and employs diffusion models to capture logical dependencies within each bipartite graph. Comparative analyses demonstrate that LayerDAG outperforms existing DAG generative models in both expressiveness and generalization, particularly for generating large-scale DAGs with up to 400 nodes—a critical scenario for system benchmarking. Extensive experiments on both synthetic and real-world flow graphs from various computing platforms show that LayerDAG generates valid DAGs with superior statistical properties and benchmarking performance. The synthetic DAGs generated by LayerDAG enhance the training of ML-based surrogate models, resulting in improved accuracy in predicting performance metrics of real-world DAGs across diverse computing platforms.

2806. AdaManip: Adaptive Articulated Object Manipulation Environments and Policy Learning

链接：https://iclr.cc/virtual/2025/poster/29969 abstract： Articulated object manipulation is a critical capability for robots to perform various tasks in real-world scenarios.Composed of multiple parts connected by joints, articulated objects are endowed with diverse functional mechanisms through complex relative motions. For example, a safe consists of a door, a handle, and a lock, where the door can only be opened when the latch is unlocked. The internal structure, such as the state of a lock or joint angle constraints, cannot be directly observed from visual observation. Consequently, successful manipulation of these objects requires adaptive adjustment based on trial and error rather than a one-time visual inference. However, previous datasets and simulation environments for articulated objects have primarily focused on simple manipulation mechanisms where the complete manipulation process can be inferred from the object's appearance. To enhance the diversity and complexity of adaptive manipulation mechanisms, we build a novel articulated object manipulation environment and equip it with 9 categories of objects. Based on the environment and objects, we further propose an adaptive demonstration collection and 3D visual diffusion-based imitation learning pipeline that learns the adaptive manipulation policy. The effectiveness of our designs and proposed method is validated through both simulation and real-world experiments.

2807. MMFakeBench: A Mixed-Source Multimodal Misinformation Detection Benchmark for LVLMs

链接：https://iclr.cc/virtual/2025/poster/30477 abstract： Current multimodal misinformation detection (MMD) methods often assume a single source and type of forgery for each sample, which is insufficient for real-world scenarios where multiple forgery sources coexist. The lack of a benchmark for mixed-source misinformation has hindered progress in this field. To address this, we introduce MMFakeBench, the first comprehensive benchmark for mixed-source MMD. MMFakeBench includes 3 critical sources: textual veracity distortion, visual veracity distortion, and cross-modal consistency distortion, along with 12 sub-categories of misinformation forgery types. We further conduct an extensive evaluation of 6 prevalent detection methods and 15 Large Vision-Language Models (LVLMs) on MMFakeBench under a zero-shot setting. The results indicate that current methods struggle under this challenging and realistic mixed-source MMD setting. Additionally, we propose MMD-Agent, a novel approach to integrate the reasoning, action, and tool-use capabilities of LVLM agents, significantly enhancing accuracy and generalization. We believe this study will catalyze future research into more realistic mixed-source multimodal misinformation and provide a fair evaluation of misinformation detection methods.

2808. Beyond Single Concept Vector: Modeling Concept Subspace in LLMs with Gaussian Distribution

链接：https://iclr.cc/virtual/2025/poster/30487 abstract： Probing learned concepts in large language models (LLMs) is crucial for understanding how semantic knowledge is encoded internally. Training linear classifiers on probing tasks is a principle approach to denote the vector of a certain concept in the representation space. However, the single vector identified for a concept varies with both data and training, making it less robust and weakening its effectiveness in real-world applications. To address this challenge, we propose an approach to approximate the subspace representing a specific concept. Built on linear probing classifiers, we extend the concept vectors into Gaussian Concept Subspace (GCS). We demonstrate GCS's effectiveness through measuring its faithfulness and plausibility across multiple LLMs with different sizes and architectures. Additionally, we use representation intervention tasks to showcase its efficacy in real-world applications such as emotion steering. Experimental results indicate that GCS concept vectors have the potential to balance steering performance and maintaining the fluency in natural language generation tasks.

2809. Non-myopic Generation of Language Models for Reasoning and Planning

链接：https://iclr.cc/virtual/2025/poster/29799 abstract： Large Language Models (LLMs) have demonstrated remarkable abilities in reasoning and planning. Despite their success in various domains, such as mathematical problem-solving and coding, LLMs face challenges in ensuring reliable and optimal planning due to the inherent myopic nature of autoregressive decoding. This paper revisits LLM reasoning from an optimal control perspective, proposing a novel method, Predictive-Decoding, that leverages Model Predictive Control to enhance planning accuracy. By reweighting LLM distributions based on foresight trajectories, Predictive-Decoding aims to mitigate early errors and promote non-myopic planning. Our experiments show significant improvements across a wide range of tasks in math, coding, and agent-based scenarios. Furthermore, Predictive-Decoding demonstrates computational efficiency, outperforming search baselines while utilizing inference compute more effectively. This study provides insights into optimizing LLM planning capabilities.

2810. On the Optimization and Generalization of Multi-head Attention

链接：https://iclr.cc/virtual/2025/poster/31501 abstract： The training and generalization dynamics of the Transformer's core mechanism, namely the Attention mechanism, remain under-explored. Besides, existing analyses primarily focus on single-head attention. Inspired by the demonstrated benefits of overparameterization when training fully-connected networks, we investigate the potential optimization and generalization advantages of using multiple attention heads. Towards this goal, we derive convergence and generalization guarantees for gradient-descent training of a single-layer multi-head self-attention model, under a suitable realizability condition on the data. We then establish primitive conditions on the initialization that ensure realizability holds. Finally, we demonstrate that these conditions are satisfied for a simple tokenized-mixture model. We expect the analysis can be extended to various data-model and architecture variations.

2811. Tamper-Resistant Safeguards for Open-Weight LLMs

链接：https://iclr.cc/virtual/2025/poster/31026 abstract： Rapid advances in the capabilities of large language models (LLMs) have raised widespread concerns regarding their potential for malicious use. Open-weight LLMs present unique challenges, as existing safeguards lack robustness to tampering attacks that modify model weights. For example, recent works have demonstrated that refusal and unlearning safeguards can be trivially removed with a few steps of fine-tuning. These vulnerabilities necessitate new approaches for enabling the safe release of open-weight LLMs. We develop a method, called TAR, for building tamper-resistant safeguards into open-weight LLMs such that adversaries cannot remove the safeguards even after hundreds of steps of fine-tuning. In extensive evaluations and red teaming analyses, we find that our method greatly improves tamper-resistance while preserving benign capabilities. Our results demonstrate that progress on tamper-resistance is possible, opening up a promising new avenue to improve the safety and security of open-weight LLMs.

2812. DCT-CryptoNets: Scaling Private Inference in the Frequency Domain

链接：https://iclr.cc/virtual/2025/poster/28520 abstract： The convergence of fully homomorphic encryption (FHE) and machine learning offers unprecedented opportunities for private inference of sensitive data. FHE enables computation directly on encrypted data, safeguarding the entire machine learning pipeline, including data and model confidentiality. However, existing FHE-based implementations for deep neural networks face significant challenges in computational cost, latency, and scalability, limiting their practical deployment. This paper introduces DCT-CryptoNets, a novel approach that operates directly in the frequency-domain to reduce the burden of computationally expensive non-linear activations and homomorphic bootstrap operations during private inference. It does so by utilizing the discrete cosine transform (DCT), commonly employed in JPEG encoding, which has inherent compatibility with remote computing services where images are generally stored and transmitted in this encoded format. DCT-CryptoNets demonstrates a substantial latency reductions of up to 5.3$\times$ compared to prior work on benchmark image classification tasks. Notably, it demonstrates inference on the ImageNet dataset within 2.5 hours (down from 12.5 hours on equivalent 96-thread compute resources). Furthermore, by *learning* perceptually salient low-frequency information DCT-CryptoNets improves the reliability of encrypted predictions compared to RGB-based networks by reducing error accumulating homomorphic bootstrap operations. DCT-CryptoNets also demonstrates superior scalability to RGB-based networks by further reducing computational cost as image size increases. This study demonstrates a promising avenue for achieving efficient and practical private inference of deep learning models on high resolution images seen in real-world applications.

2813. Learning Structured Representations by Embedding Class Hierarchy with Fast Optimal Transport

链接：https://iclr.cc/virtual/2025/poster/30615 abstract： To embed structured knowledge within labels into feature representations, prior work (Zeng et al., 2022) proposed to use the Cophenetic Correlation Coefficient (CPCC) as a regularizer during supervised learning. This regularizer calculates pairwise Euclidean distances of class means and aligns them with the corresponding shortest path distances derived from the label hierarchy tree. However, class means may not be good representatives of the class conditional distributions, especially when they are multi-mode in nature. To address this limitation, under the CPCC framework, we propose to use the Earth Mover's Distance (EMD) to measure the pairwise distances among

classes in the feature space. We show that our exact EMD method generalizes previous work, and recovers the existing algorithm when class-conditional distributions are Gaussian in the feature space. To further improve the computational efficiency of our method, we introduce the Optimal Transport-CPCC family by exploring four EMD approximation variants. Our most efficient OT-CPCC variant runs in linear time in the size of the dataset, while maintaining competitive performance across datasets and tasks. The code is available at https://github.com/uiuctml/OTCPCC.

2814. Words in Motion: Extracting Interpretable Control Vectors for Motion Transformers

链接：https://iclr.cc/virtual/2025/poster/30124 abstract： Transformer-based models generate hidden states that are difficult to interpret. In this work, we analyze hidden states and modify them at inference, with a focus on motion forecasting. We use linear probing to analyze whether interpretable features are embedded in hidden states. Our experiments reveal high probing accuracy, indicating latent space regularities with functionally important directions. Building on this, we use the directions between hidden states with opposing features to fit control vectors. At inference, we add our control vectors to hidden states and evaluate their impact on predictions. Remarkably, such modifications preserve the feasibility of predictions. We further refine our control vectors using sparse autoencoders (SAEs). This leads to more linear changes in predictions when scaling control vectors. Our approach enables mechanistic interpretation as well as zero-shot generalization to unseen dataset characteristics with negligible computational overhead.

2815. FOSP: Fine-tuning Offline Safe Policy through World Models

链接：https://iclr.cc/virtual/2025/poster/28976 abstract： Offline Safe Reinforcement Learning (RL) seeks to address safety constraints by learning from static datasets and restricting exploration. However, these approaches heavily rely on the dataset and struggle to generalize to unseen scenarios safely. In this paper, we aim to improve safety during the deployment of vision-based robotic tasks through online fine-tuning an offline pretrained policy. To facilitate effective fine-tuning, we introduce model-based RL, which is known for its data efficiency. Specifically, our method employs in-sample optimization to improve offline training efficiency while incorporating reachability guidance to ensure safety. After obtaining an offline safe policy, a safe policy expansion approach is leveraged for online fine-tuning. The performance of our method is validated on simulation benchmarks with five vision-only tasks and through real-world robot deployment using limited data. It demonstrates that our approach significantly improves the generalization of offline policies to unseen safety-constrained scenarios. To the best of our knowledge, this is the first work to explore offline-to-online RL for safe generalization tasks. The videos are available at https://sunlighted.github.io/fosp_web/.

2816. Adversaries With Incentives: A Strategic Alternative to Adversarial Robustness

链接：https://iclr.cc/virtual/2025/poster/29487 abstract： Adversarial training aims to defend against adversaries: malicious opponents whose sole aim is to harm predictive performance in any way possible. This presents a rather harsh perspective, which we assert results in unnecessarily conservative training. As an alternative, we propose to model opponents as simply pursuing their own goals—rather than working directly against the classifier. Employing tools from strategic modeling, our approach enables knowledge or beliefs regarding the opponent's possible incentives to be used as inductive bias for learning. Accordingly, our method of strategic training is designed to defend against all opponents within an `incentive uncertainty set'. This resorts to adversarial learning when the set is maximal, but offers potential gains when the set can be appropriately reduced. We conduct a series of experiments that show how even mild knowledge regarding the opponent's incentives can be useful, and that the degree of potential gains depends on how these incentives relate to the structure of the learning task.

2817. Ctrl-U: Robust Conditional Image Generation via Uncertainty-aware Reward Modeling

链接：https://iclr.cc/virtual/2025/poster/28943 abstract： In this paper, we focus on the task of conditional image generation, where an image is synthesized according to user instructions. The critical challenge underpinning this task is ensuring both the fidelity of the generated images and their semantic alignment with the provided conditions. To tackle this issue, previous studies have employed supervised perceptual losses derived from pre-trained models, i.e., reward models, to enforce alignment between the condition and the generated result. However, we observe one inherent shortcoming: considering the diversity of synthesized images, the reward model usually provides inaccurate feedback when encountering newly generated data, which can undermine the training process. To address this limitation, we propose an uncertainty-aware reward modeling, called Ctrl-U, including uncertainty estimation and uncertainty-aware regularization, designed to reduce the adverse effects of imprecise feedback from the reward model. Given the inherent cognitive uncertainty within reward models, even images generated under identical conditions often result in a relatively large discrepancy in reward loss. Inspired by the observation, we explicitly leverage such prediction variance as an uncertainty indicator. Based on the uncertainty estimation, we regularize the model training by adaptively rectifying the reward. In particular, rewards with lower uncertainty receive higher loss weights, while those with higher uncertainty are given reduced weights to allow for larger variability. The proposed uncertainty regularization facilitates reward fine-tuning through consistency construction. Extensive experiments validate the effectiveness of our methodology in improving the controllability and generation quality, as well as its scalability across diverse conditional scenarios, including segmentation

mask, edge, and depth conditions.

2818. SAGEPhos: Sage Bio-Coupled and Augmented Fusion for Phosphorylation Site Detection

链接：https://iclr.cc/virtual/2025/poster/28768 abstract： Phosphorylation site prediction based on kinase-substrate interaction plays a vital role in understanding cellular signaling pathways and disease mechanisms. Computational methods for this task can be categorized into kinase-family-focused and individual kinase-targeted approaches. Individual kinase-targeted methods have gained prominence for their ability to explore a broader protein space and provide more precise target information for kinase inhibitors. However, most existing individual kinase-based approaches focus solely on sequence inputs, neglecting crucial structural information. To address this limitation, we introduce SAGEPhos (Structure-aware kinAse-substrate bio-coupled and bio-auGmented nEtwork for Phosphorylation site prediction), a novel framework that modifies the semantic space of main protein inputs using auxiliary inputs at two distinct modality levels. At the inter-modality level, SAGEPhos introduces a Bio-Coupled Modal Fusion method, distilling essential kinase sequence information to refine task-oriented local substrate feature space, creating a shared semantic space that captures crucial kinase-substrate interaction patterns. Within the substrate's intra-modality domain, it focuses on Bio-Augmented Fusion, emphasizing 2D local sequence information while selectively incorporating 3D spatial information from predicted structures to complement the sequence space. Moreover, to address the lack of structural information in current datasets, we contribute a new, refined phosphorylation site prediction dataset, which incorporates crucial structural elements and will serve as a new benchmark for the field. Experimental results demonstrate that SAGEPhos significantly outperforms baseline methods, notably achieving almost 10\% and 12\% improvements in prediction accuracy and AUC-ROC, respectively. We further demonstrate our algorithm's robustness and generalization through stable results across varied data partitions and significant improvements in zero-shot scenarios. These results underscore the effectiveness of constructing a larger and more precise protein space in advancing the state-of-the-art in phosphorylation site prediction. We release the SAGEPhos models and code at https://github.com/ZhangJJ26/SAGEPhos.

2819. MaRS: A Fast Sampler for Mean Reverting Diffusion based on ODE and SDE Solvers

链接：https://iclr.cc/virtual/2025/poster/27721 abstract： In applications of diffusion models, controllable generation is of practical significance, but is also challenging. Current methods for controllable generation primarily focus on modifying the score function of diffusion models, while Mean Reverting (MR) Diffusion directly modifies the structure of the stochastic differential equation (SDE), making the incorporation of image conditions simpler and more natural. However, current training-free fast samplers are not directly applicable to MR Diffusion. And thus MR Diffusion requires hundreds of NFEs (number of function evaluations) to obtain high-quality samples. In this paper, we propose a new algorithm named MaRS (MR Sampler) to reduce the sampling NFEs of MR Diffusion. We solve the reverse-time SDE and the probability flow ordinary differential equation (PF-ODE) associated with MR Diffusion, and derive semi-analytical solutions. The solutions consist of an analytical function and an integral parameterized by a neural network. Based on this solution, we can generate high-quality samples in fewer steps. Our approach does not require training and supports all mainstream parameterizations, including noise prediction, data prediction and velocity prediction. Extensive experiments demonstrate that MR Sampler maintains high sampling quality with a speedup of 10 to 20 times across ten different image restoration tasks. Our algorithm accelerates the sampling procedure of MR Diffusion, making it more practical in controllable generation.

2820. Precise Localization of Memories: A Fine-grained Neuron-level Knowledge Editing Technique for LLMs

链接：https://iclr.cc/virtual/2025/poster/30923 abstract： Knowledge editing aims to update outdated information in Large Language Models (LLMs). A representative line of study is locate-then-edit methods, which typically employ causal tracing to identify the modules responsible for recalling factual knowledge about entities. However, we find these methods are often sensitive only to changes in the subject entity, leaving them less effective at adapting to changes in relations. This limitation results in poor editing locality, which can lead to the persistence of irrelevant or inaccurate facts, ultimately compromising the reliability of LLMs. We believe this issue arises from the insufficient precision of knowledge localization. To address this, we propose a Fine-grained Neuron-level Knowledge Editing (FiNE) method that enhances editing locality without affecting overall success rates. By precisely identifying and modifying specific neurons within feed-forward networks, FiNE significantly improves knowledge localization and editing. Quantitative experiments demonstrate that FiNE efficiently achieves better overall performance compared to existing techniques, providing new insights into the localization and modification of knowledge within LLMs.

2821. CatVTON: Concatenation Is All You Need for Virtual Try-On with Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/28612 abstract： Virtual try-on methods based on diffusion models achieve realistic effects but often require additional encoding modules, a large number of training parameters, and complex preprocessing, which increases the burden on training and inference. In this work, we re-evaluate the necessity of additional modules and analyze how to improve training efficiency and reduce redundant steps in the inference process. Based on these insights, we propose

CatVTON, a simple and efficient virtual try-on diffusion model that transfers in-shop or worn garments of arbitrary categories to target individuals by concatenating them along spatial dimensions as inputs of the diffusion model. The efficiency of CatVTON is reflected in three aspects: (1) Lightweight network. CatVTON consists only of a VAE and a simplified denoising UNet, removing redundant image and text encoders as well as cross-attentions, and includes just 899.06M parameters. (2) Parameter-efficient training. Through experimental analysis, we identify self-attention modules as crucial for adapting pre-trained diffusion models to the virtual try-on task, enabling high-quality results with only 49.57M training parameters. (3) Simplified inference. CatVTON eliminates unnecessary preprocessing, such as pose estimation, human parsing, and captioning, requiring only a person image and garment reference to guide the virtual try-on process, reducing over 49% memory usage compared to other diffusion-based methods. Extensive experiments demonstrate that CatVTON achieves superior qualitative and quantitative results compared to baseline methods and demonstrates strong generalization performance in in-the-wild scenarios, despite being trained solely on public datasets with 73K samples.

2822. Progressive Mixed-Precision Decoding for Efficient LLM Inference

链接：https://iclr.cc/virtual/2025/poster/29823 abstract： In spite of the great potential of large language models (LLMs) across various tasks, their deployment on resource-constrained devices remains challenging due to their excessive computational and memory demands. Quantization has emerged as an effective solution by storing weights in reduced precision. However, utilizing low precisions (i.e.~2/3-bit) to substantially alleviate the memory-boundedness of LLM decoding, still suffers from prohibitive performance drop. In this work, we argue that existing approaches fail to explore the diversity in computational patterns, redundancy, and sensitivity to approximations of the different phases of LLM inference, resorting to a uniform quantization policy throughout.Instead, we propose a novel phase-aware method that selectively allocates precision during different phases of LLM inference, achieving both strong context extraction during prefill and efficient memory bandwidth utilization during decoding. To further address the memory-boundedness of the decoding phase, we introduce Progressive Mixed-Precision Decoding (PMPD), a technique that enables the gradual lowering of precision deeper in the generated sequence, together with a spectrum of precision-switching schedulers that dynamically drive the precision-lowering decisions in either task-adaptive or prompt-adaptive manner. Extensive evaluation across diverse language tasks shows that when targeting Nvidia GPUs, PMPD achieves 1.4$-$12.2$\times$ speedup in matrix-vector multiplications over fp16 models, while when targeting an LLM-optimized NPU, our approach delivers a throughput gain of 3.8$-$8.0$\times$ over fp16 models and up to 1.54$\times$ over uniform quantization approaches while preserving the output quality.

2823. SC-OmniGS: Self-Calibrating Omnidirectional Gaussian Splatting

链接：https://iclr.cc/virtual/2025/poster/30809 abstract： 360-degree cameras streamline data collection for radiance field 3D reconstruction by capturing comprehensive scene data. However, traditional radiance field methods do not address the specific challenges inherent to 360-degree images. We present SC-OmniGS, a novel self-calibrating omnidirectional Gaussian splatting system for fast and accurate omnidirectional radiance field reconstruction using 360-degree images. Rather than converting 360-degree images to cube maps and performing perspective image calibration, we treat 360-degree images as a whole sphere and derive a mathematical framework that enables direct omnidirectional camera pose calibration accompanied by 3D Gaussians optimization. Furthermore, we introduce a differentiable omnidirectional camera model in order to rectify the distortion of real-world data for performance enhancement. Overall, the omnidirectional camera intrinsic model, extrinsic poses, and 3D Gaussians are jointly optimized by minimizing weighted spherical photometric loss. Extensive experiments have demonstrated that our proposed SC-OmniGS is able to recover a high-quality radiance field from noisy camera poses or even no pose prior in challenging scenarios characterized by wide baselines and non-object-centric configurations. The noticeable performance gain in the real-world dataset captured by consumer-grade omnidirectional cameras verifies the effectiveness of our general omnidirectional camera model in reducing the distortion of 360-degree images.

2824. DRESSing Up LLM: Efficient Stylized Question-Answering via Style Subspace Editing

链接：https://iclr.cc/virtual/2025/poster/28469 abstract： We introduce DRESS, a novel approach for generating stylized large language model (LLM) responses through representation editing. Existing methods like prompting and fine-tuning are either insufficient for complex style adaptation or computationally expensive, particularly in tasks like NPC creation or character role-playing. Our approach leverages the over-parameterized nature of LLMs to disentangle a style-relevant subspace within the model's representation space to conduct representation editing, ensuring a minimal impact on the original semantics. By applying adaptive editing strengths, we dynamically adjust the steering vectors in the style subspace to maintain both stylistic fidelity and semantic integrity. We develop two stylized QA benchmark datasets to validate the effectiveness of DRESS, and the results demonstrate significant improvements compared to baseline methods such as prompting and ITI. In short, DRESS is a lightweight, train-free solution for enhancing LLMs with flexible and effective style control, making it particularly useful for developing stylized conversational agents. Codes and benchmark datasets are available at https://github.com/ArthurLeoM/DRESS-LLM.

2825. EmbedLLM: Learning Compact Representations of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30322 abstract： With hundreds of thousands of language models available on

Huggingface today, efficiently evaluating and utilizing these models across various downstream tasks has become increasingly critical. Many existing methods repeatedly learn task-specific representations of Large Language Models (LLMs), which leads to inefficiencies in both time and computational resources. To address this, we propose EmbedLLM, a framework designed to learn compact vector representations of LLMs that facilitate downstream applications involving many models, such as model routing. We introduce an encoder-decoder approach for learning such embedding, along with a systematic framework to evaluate their effectiveness. Empirical results show that EmbedLLM outperforms prior methods in model routing. Additionally, we demonstrate that our method can forecast a model's performance on multiple benchmarks, without incurring additional inference cost. Extensive probing experiments validate that the learned embeddings capture key model characteristics, e.g. whether the model is specialized for coding tasks, even without being explicitly trained on them. We open source our dataset, code and embedder to facilitate further research and application.

2826. AgentSquare: Automatic LLM Agent Search in Modular Design Space

链接：https://iclr.cc/virtual/2025/poster/32059 abstract： Recent advancements in Large Language Models (LLMs) have led to a rapid growth of agentic systems capable of handling a wide range of complex tasks. However, current research largely relies on manual, task-specific design, limiting their adaptability to novel tasks. In this paper, we introduce a new research problem: Modularized LLM Agent Search (MoLAS). We propose a modular design space that abstracts existing LLM agent designs into four fundamental modules with uniform IO interface: Planning, Reasoning, Tool Use, and Memory. Building on this design space, we present a novel LLM agent search framework called AgentSquare, which introduces two core mechanisms, i.e., module evolution and recombination, to efficiently search for optimized LLM agents. To further accelerate the process, we design a performance predictor that uses in-context surrogate models to skip unpromising agent designs. Extensive experiments across six benchmarks, covering the diverse scenarios of web, embodied, tool use and game applications, show that AgentSquare substantially outperforms hand-crafted agents, achieving an average performance gain of 17.2% against best-known human designs. Moreover, AgentSquare can generate interpretable design insights, enabling a deeper understanding of agentic architecture and its impact on task performance. We believe that the modular design space and AgentSquare search framework offer a platform for fully exploiting the potential of prior successful designs and consolidate the collective efforts of research community. Code repo is available at https://github.com/tsinghua-fib-lab/AgentSquare.

2827. ActSafe: Active Exploration with Safety Constraints for Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29177 abstract： Reinforcement learning (RL) is ubiquitous in the development of modern AI systems. However, state-of-the-art RL agents require extensive, and potentiallyunsafe, interactions with their environments to learn effectively. These limitationsconfine RL agents to simulated environments, hindering their ability to learndirectly in real-world settings. In this work, we present ActSafe, a novelmodel-based RL algorithm for safe and efficient exploration. ActSafe learnsa well-calibrated probabilistic model of the system and plans optimisticallyw.r.t. the epistemic uncertainty about the unknown dynamics, while enforcingpessimism w.r.t. the safety constraints. Under regularity assumptions on theconstraints and dynamics, we show that ActSafe guarantees safety duringlearning while also obtaining a near-optimal policy in finite time. In addition, wepropose a practical variant of ActSafe that builds on latest model-based RL advancements and enables safe exploration even in high-dimensional settings suchas visual control. We empirically show that ActSafe obtains state-of-the-artperformance in difficult exploration tasks on standard safe deep RL benchmarkswhile ensuring safety during learning.

2828. TPO: Aligning Large Language Models with Multi-branch & Multi-step Preference Trees

链接：https://iclr.cc/virtual/2025/poster/29846 abstract： In the domain of complex reasoning tasks, such as mathematical reasoning, recent advancements have proposed the use of Direct Preference Optimization (DPO) to suppress output of dispreferred responses, thereby enhancing the long-chain reasoning capabilities of large language models (LLMs). To this end, these studies employed LLMs to generate preference trees via Tree-of-thoughts (ToT) and sample the paired preference responses required by the DPO algorithm. However, the DPO algorithm based on binary preference optimization is unable to learn multiple responses with varying degrees of preference/dispreference that provided by the preference trees, resulting in incomplete preference learning. In this work, we introduce Tree Preference Optimization (TPO), that does not sample paired preference responses from the preference tree; instead, it directly learns from the entire preference tree during the fine-tuning. Specifically, TPO formulates the language model alignment as a Preference List Ranking problem, where the policy can potentially learn more effectively from a ranked preference list of responses given the prompt. In addition, to further assist LLMs in identifying discriminative steps within long-chain reasoning and increase the relative reward margin in the preference list, TPO utilizes Adaptive Step Reward to adjust the reward values of each step in trajectory for performing fine-grained preference optimization. We carry out extensive experiments on mathematical reasoning tasks to evaluate TPO. The experimental results indicate that TPO consistently outperforms DPO across five public large language models on four datasets.

2829. Bridging the Gap Between f-divergences and Bayes Hilbert Spaces

链接：https://iclr.cc/virtual/2025/poster/28483 abstract： We introduce a novel framework that generalizes f-divergences by incorporating locally non-convex divergence-generating functions.Using this extension, we define a new class of pseudo f-

divergences, encompassing a wider range of distances between distributions that traditional f-divergences cannot capture.Among these, we focus on a particular pseudo divergence obtained by considering the induced metric of Bayes Hilbert spaces.Bayes Hilbert spaces are frequently used due to their inherent connection to Bayes's theorem. They allow sampling from potentially intractable posterior densities, which has remained challenging until now.In the more general context, we prove that pseudo f-divergences are well-defined and introduce a variational estimation framework that can be used in a statistical learning context.By applying this variational estimation framework to f-GANs, we achieve improved FID scores over existing f-GAN architectures and competitive results with the Wasserstein GAN, highlighting its potential for both theoretical research and practical applications in learning theory.

2830. Designing Mechanical Meta-Materials by Learning Equivariant Flows

链接：https://iclr.cc/virtual/2025/poster/29419 abstract： Mechanical meta-materials are solids whose geometric structure results in exotic nonlinear behaviors that are not typically achievable via homogeneous materials. We show how to drastically expand the design space of a class of mechanical meta-materials known as $\textit{cellular solids}$, by generalizing beyond translational symmetry. This is made possible by transforming a reference geometry according to a divergence free flow that is parameterized by a neural network and equivariant under the relevant symmetry group. We show how to construct flows equivariant to the space groups, despite the fact that these groups are not compact. Coupling this flow with a differentiable nonlinear mechanics simulator allows us to represent a much richer set of cellular solids than was previously possible. These materials can be optimized to exhibit desirable mechanical properties such as negative Poisson's ratios or to match target stress-strain curves. We validate these new designs in simulation and by fabricating real-world prototypes. We find that designs with higher-order symmetries can exhibit a wider range of behaviors.

2831. DeepTAGE: Deep Temporal-Aligned Gradient Enhancement for Optimizing Spiking Neural Networks

链接：https://iclr.cc/virtual/2025/poster/28961 abstract： Spiking Neural Networks (SNNs), with their biologically inspired spatio-temporal dynamics and spike-driven processing, are emerging as a promising low-power alternative to traditional Artificial Neural Networks (ANNs). However, the complex neuronal dynamics and non-differentiable spike communication mechanisms in SNNs present substantial challenges for efficient training. By analyzing the membrane potentials in spiking neurons, we found that their distributions can increasingly deviate from the firing threshold as time progresses, which tends to cause diminished backpropagation gradients and unbalanced optimization. To address these challenges, we propose Deep Temporal-Aligned Gradient Enhancement (DeepTAGE), a novel approach that improves optimization gradients in SNNs from both internal surrogate gradient functions and external supervision methods. Our DeepTAGE dynamically adjusts surrogate gradients in accordance with the membrane potential distribution across different time steps, enhancing their respective gradients in a temporal-aligned manner that promotes balanced training. Moreover, to mitigate issues of gradient vanishing or deviating during backpropagation, DeepTAGE incorporates deep supervision at both spatial (network stages) and temporal (time steps) levels to ensure more effective and robust network optimization. Importantly, our method can be seamlessly integrated into existing SNN architectures without imposing additional inference costs or requiring extra control modules. We validate the efficacy of DeepTAGE through extensive experiments on static benchmarks (CIFAR10, CIFAR100, and ImageNet-1k) and a neuromorphic dataset (DVS-CIFAR10), demonstrating significant performance improvements.

2832. Round and Round We Go! What makes Rotary Positional Encodings useful?

链接：https://iclr.cc/virtual/2025/poster/30251 abstract： Positional Encodings (PEs) are a critical component of Transformer-based Large Language Models (LLMs), providing the attention mechanism with important sequence-position information. One of the most popular types of encoding used today in LLMs are Rotary Positional Encodings (RoPE), that rotate the queries and keys based on their relative distance. A common belief is that RoPE is useful because it helps to decay token dependency as relative distance increases. In this work, we argue that this is unlikely to be the core reason. We study the internals of a trained Gemma 7B model to understand how RoPE is being used at a mechanical level. We find that Gemma learns to use RoPE to construct robust `positional' attention patterns by exploiting the highest frequencies. We also find that, in general, Gemma greatly prefers to use the lowest frequencies of RoPE, which we suspect are used to carry semantic information. We mathematically prove interesting behaviours of RoPE and conduct experiments to verify our findings, proposing a modification of RoPE that fixes some highlighted issues and improves performance. We believe that this work represents an interesting step in better understanding PEs in LLMs, which we believe holds crucial value for scaling LLMs to large sizes and context lengths.

2833. SysCaps: Language Interfaces for Simulation Surrogates of Complex Systems

链接：https://iclr.cc/virtual/2025/poster/31201 abstract： Surrogate models are used to predict the behavior of complex energy systems that are too expensive to simulate with traditional numerical methods. Our work introduces the use of language descriptions, which we call "system captions" or SysCaps, to interface with such surrogates. We argue that interacting with surrogates through text, particularly natural language, makes these models more accessible for both experts and non-experts.We introduce a lightweight multimodal text and timeseries regression model and a training pipeline that uses large

language models (LLMs) to synthesize high-quality captions from simulation metadata. Our experiments on two real-world simulators of buildings and wind farms show that our SysCaps-augmented surrogates have better accuracy on held-out systems than traditional methods while enjoying new generalization abilities, such as handling semantically related descriptions of the same test system.Additional experiments also highlight the potential of SysCaps to unlock language-driven design space exploration and to regularize training through prompt augmentation.

2834. Sparse Learning for State Space Models on Mobile

链接：https://iclr.cc/virtual/2025/poster/28076 abstract： Transformer models have been widely investigated in different domains by providing long-range dependency handling and global contextual awareness, driving the development of popular AI applications such as ChatGPT, Gemini, and Alexa.State Space Models (SSMs) have emerged as strong contenders in the field of sequential modeling, challenging the dominance of Transformers. SSMs incorporate a selective mechanism that allows for dynamic parameter adjustment based on input data, enhancing their performance.However, this mechanism also comes with increasing computational complexity and bandwidth demands, posing challenges for deployment on resource-constraint mobile devices.To address these challenges without sacrificing the accuracy of the selective mechanism, we propose a sparse learning framework that integrates architecture-aware compiler optimizations. We introduce an end-to-end solution--$\mathbf{C}_4^n$ kernel sparsity, which prunes n elements from every four contiguous weights, and develop a compiler-based acceleration solution to ensure execution efficiency for this sparsity on mobile devices.Based on the kernel sparsity, our framework generates optimized sparse models targeting specific sparsity or latency requirements for various model sizes. We further leverage pruned weights to compensate for the remaining weights, enhancing downstream task performance.For practical hardware acceleration, we propose $\mathbf{C}_4^n$-specific optimizations combined with a layout transformation elimination strategy. This approach mitigates inefficiencies arising from fine-grained pruning in linear layers and improves performance across other operations. Experimental results demonstrate that our method achieves superior task performance compared to other semi-structured pruning methods and achieves up-to 7$\times$ speedup compared to llama.cpp framework on mobile devices.

2835. Revisit the Open Nature of Open Vocabulary Semantic Segmentation

链接：https://iclr.cc/virtual/2025/poster/31104 abstract： In Open Vocabulary Semantic Segmentation (OVS), we observe a consistent dropin model performance as the query vocabulary set expands, especially when itincludes semantically similar and ambiguous vocabularies, such as ‘sofa’ and‘couch’. The previous OVS evaluation protocol, however, does not account forsuch ambiguity, as any mismatch between model-predicted and human-annotatedpairs is simply treated as incorrect on a pixel-wise basis. This contradicts the opennature of OVS, where ambiguous categories may both be correct from an open-world perspective. To address this, in this work, we study the open nature of OVSand propose a mask-wise evaluation protocol that is based on matched and mis-matched mask pairs between prediction and annotation respectively. Extensiveexperimental evaluations show that the proposed mask-wise protocol provides amore effective and reliable evaluation framework for OVS models compared to theprevious pixel-wise approach on the perspective of open-world. Moreover, analy-sis of mismatched mask pairs reveals that a large amount of ambiguous categoriesexist in commonly used OVS datasets. Interestingly, we find that reducing theseambiguities during both training and inference enhances capabilities of OVS mod-els. These findings and the new evaluation protocol encourage further explorationof the open nature of OVS, as well as broader open-world challenges. Project page: https://qiming-huang.github.io/RevisitOVS/.

2836. Are Transformers Able to Reason by Connecting Separated Knowledge in Training Data?

链接：https://iclr.cc/virtual/2025/poster/31196 abstract： Humans exhibit remarkable compositional reasoning by integrating knowledge from various sources. For example, if someone learns (B = f(A)) from one source and (C = g(B)) from another, they can deduce (C=g(B)=g(f(A))) even without encountering (ABC) together, showcasing the generalization ability of human intelligence. In this paper, we introduce a synthetic learning task, "FTCT" (Fragmented at Training, Chained at Testing), to validate the potential of Transformers in replicating this skill and interpret its inner mechanism. During training, data consist of separated knowledge fragments from an overall causal graph. In testing, Transformers must combine these fragments to infer complete causal traces. Our findings demonstrate that few-shot Chain-of-Thought prompting enables Transformers to perform compositional reasoning on FTCT by revealing correct combinations of fragments, even if such combinations were absent in training data. Furthermore, the emergence of compositional reasoning ability is strongly correlated with model complexity and training-testing data similarity. We propose, both theoretically and empirically, that Transformers learn an underlying generalizable program from training, enabling effective compositional reasoning during testing.

2837. Multi-Scale Fusion for Object Representation

链接：https://iclr.cc/virtual/2025/poster/28394 abstract： Representing images or videos as object-level feature vectors, rather than pixel-level feature maps, facilitates advanced visual tasks.Object-Centric Learning (OCL) primarily achieves this by reconstructing the input under the guidance of Variational Autoencoder (VAE) intermediate representation to drive so-called slots to aggregate as much object information as possible.However, existing VAE guidance does not explicitly address that objects can vary in pixel sizes while models typically excel at specific pattern scales.We propose Multi-Scale Fusion (MSF) to enhance VAE guidance for OCL training. To ensure objects of all sizes fall within VAE's comfort zone, we adopt the image pyramid, which produces intermediate representations at multiple scales;To foster scale-invariance/variance in object super-pixels, we devise inter/intra-scale fusion, which augments low-quality object super-pixels of one scale with corresponding high-

quality super-pixels from another scale.On standard OCL benchmarks, our technique improves mainstream methods, including state-of-the-art diffusion-based ones.The source code is available on https://github.com/Genera1Z/MultiScaleFusion.

2838. Recognize Any Surgical Object: Unleashing the Power of Weakly-Supervised Data

链接：https://iclr.cc/virtual/2025/poster/28666 abstract： We present RASO, a foundation model designed to Recognize Any Surgical Object, offering robust open-set recognition capabilities across a broad range of surgical procedures and object classes, in both surgical images and videos. RASO leverages a novel weakly-supervised learning framework that generates tag-image-text pairs automatically from large-scale unannotated surgical lecture videos, significantly reducing the need for manual annotations. Our scalable data generation pipeline gathers 2,200 surgical procedures and produces 3.6 million tag annotations across 2,066 unique surgical tags. Our experiments show that RASO achieves improvements of 2.9 mAP, 4.5 mAP, 10.6 mAP, and 7.2 mAP on four standard surgical benchmarks respectively in zero-shot settings, and surpasses state-of-the-art models in supervised surgical action recognition tasks. We will open-source our code, model, and dataset to facilitate further research.

2839. GOLD: Graph Out-of-Distribution Detection via Implicit Adversarial Latent Generation

链接：https://iclr.cc/virtual/2025/poster/27751 abstract： Despite graph neural networks' (GNNs) great success in modelling graph-structured data, out-of-distribution (OOD) test instances still pose a great challenge for current GNNs. One of the most effective techniques to detect OOD nodes is to expose the detector model with an additional OOD node-set, yet the extra OOD instances are often difficult to obtain in practice. Recent methods for image data address this problem using OOD data synthesis, typically relying on pre-trained generative models like Stable Diffusion. However, these approaches require vast amounts of additional data, as well as one-for-all pre-trained generative models, which are not available for graph data. Therefore, we propose the GOLD framework for graph OOD detection, an implicit adversarial learning pipeline with synthetic OOD exposure without pre-trained models. The implicit adversarial training process employs a novel alternating optimisation framework by training: (1) a latent generative model to regularly imitate the in-distribution (ID) embeddings from an evolving GNN, and (2) a GNN encoder and an OOD detector to accurately classify ID data while increasing the energy divergence between the ID embeddings and the generative model's synthetic embeddings. This novel approach implicitly transforms the synthetic embeddings into pseudo-OOD instances relative to the ID data, effectively simulating exposure to OOD scenarios without auxiliary data. Extensive OOD detection experiments are conducted on five benchmark graph datasets, verifying the superior performance of GOLD without using real OOD data compared with the state-of-the-art OOD exposure and non-exposure baselines.

2840. SafeWatch: An Efficient Safety-Policy Following Video Guardrail Model with Transparent Explanations

链接：https://iclr.cc/virtual/2025/poster/32048 abstract： With the rise of generative AI and rapid growth of high-quality video generation, video guardrails have become more crucial than ever to ensure safety and security across platforms. Current video guardrails, however, are either overly simplistic, relying on pure classification models trained on simple policies with limited unsafe categories, which lack detailed explanations, or prompting multimodal large language models (MLLMs) with long safety guidelines, which are inefficient and impractical for guardrailing real-world content. To bridge this gap, we propose SafeWatch, an efficient MLLM-based video guardrail model designed to follow customized safety policies and provide multi-label video guardrail outputs with content-specific explanations in a zero-shot manner. In particular, unlike traditional MLLM-based guardrails that encode all safety policies autoregressively, causing inefficiency and bias, SafeWatch uniquely encodes each policy chunk in parallel and eliminates their position bias such that all policies are attended simultaneously with equal importance. In addition, to improve efficiency and accuracy, SafeWatch incorporates a policy-aware visual token pruning algorithm that adaptively selects the most relevant video tokens for each policy, discarding noisy or irrelevant information. This allows for more focused, policy-compliant guardrail with significantly reduced computational overhead. Considering the limitations of existing video guardrail benchmarks, we propose SafeWatch-Bench, a large-scale video guardrail benchmark comprising over 2M videos spanning six safety categories which covers over 30 tasks to ensure a comprehensive coverage of all potential safety scenarios. We have conducted extensive experiments, showing that SafeWatch outperforms all SOTA video guardrails on SafeWatch-Bench by 28.2%, and achieves a 13.6% improvement on existing benchmarks, all while reducing inference costs by an average of 10%. SafeWatch also demonstrates strong policy-following abilities and outperforms previous SOTAs by 5.6% and 15.6% in zero-shot generalizability to new policies and new prompting tasks. Additionally, both LLM-as-a-judge and human evaluators confirm the high quality of the explanations provided by SafeWatch. Our project is open-sourced at https://safewatch-aiguard.github.io.

2841. Fine-tuning can cripple your foundation model; preserving features may be the solution

链接：https://iclr.cc/virtual/2025/poster/31485 abstract： Pre-trained foundation models, due to their enormous capacity and exposure to vast amounts of data during pre-training, are known to have learned plenty of real-world concepts. An important step in making these pre-trained models effective on downstream tasks is to fine-tune them on related datasets. While various fine-

tuning methods have been devised and have been shown to be highly effective, we observe that a fine-tuned model's ability to recognize concepts on tasks different from the downstream one is reduced significantly compared to its pre-trained counterpart. This is an undesirable effect of fine-tuning as a substantial amount of resources was used to learn these pre-trained concepts in the first place. We call this phenomenon "concept forgetting" and via experiments show that most end-to-end fine-tuning approaches suffer heavily from this side effect. To this end, we propose a simple fix to this problem by designing a new fine-tuning method called LDIFS (short for ℓ_2 distance in feature space) that, while learning new concepts related to the downstream task, allows a model to preserve its pre-trained knowledge as well. Through extensive experiments on 10 fine-tuning tasks we show that LDIFS significantly reduces concept forgetting. Additionally, we show that LDIFS is highly effective in performing continual fine-tuning on a sequence of tasks as well, in comparison with both fine-tuning as well as continual learning baselines.

2842. GALA: Geometry-Aware Local Adaptive Grids for Detailed 3D Generation

链接：https://iclr.cc/virtual/2025/poster/30050 abstract： We propose GALA, a novel representation of 3D shapes that (i) excels at capturing and reproducing complex geometry and surface details, (ii) is computationally efficient, and (iii) lends itself to 3D generative modelling with modern, diffusion-based schemes. The key idea of GALA is to exploit both the global sparsity of surfaces within a 3D volume and their local surface properties. Sparsity is promoted by covering only the 3D object boundaries, not empty space, with an ensemble of tree root voxels. Each voxel contains an octree to further limit storage and compute to regions that contain surfaces. Adaptivity is achieved by fitting one local and geometry-aware coordinate frame in each non-empty leaf node. Adjusting the orientation of the local grid, as well as the anisotropic scales of its axes, to the local surface shape greatly increases the amount of detail that can be stored in a given amount of memory, which in turn allows for quantization without loss of quality. With our optimized C++/CUDA implementation, GALA can be fitted to an object in less than 10 seconds. Moreover, the representation can efficiently be flattened and manipulated with transformer networks. We provide a cascaded generation pipeline capable of generating 3D shapes with great geometric detail. For more information, please visit our project page.

2843. Instructional Segment Embedding: Improving LLM Safety with Instruction Hierarchy

链接：https://iclr.cc/virtual/2025/poster/28101 abstract： Large Language Models (LLMs) are susceptible to security and safety threats, such as prompt injection, prompt extraction, and harmful requests.One major cause of these vulnerabilities is the lack of an instruction hierarchy.Modern LLM architectures treat all inputs equally, failing to distinguish between and prioritize various types of instructions, such as system messages, user prompts, and data. As a result, lower-priority user prompts may override more critical system instructions, including safety protocols. Existing approaches to achieving instruction hierarchy, such as delimiters and instruction-based training, do not address this issue at the architectural level.We introduce the $\textbf{I}$nstructional $\textbf{S}$egment $\textbf{E}$mbedding (ISE) technique, inspired by BERT, to modern large language models, which embeds instruction priority information directly into the model. This approach enables models to explicitly differentiate and prioritize various instruction types, significantly improving safety against malicious prompts that attempt to override priority rules. Our experiments on the Structured Query and Instruction Hierarchy benchmarks demonstrate an average robust accuracy increase of up to 15.75\% and 18.68\%, respectively. Furthermore, we observe an improvement in the instruction-following capability of up to 4.1\% on AlpacaEval. Overall, our approach offers a promising direction for enhancing the safety and effectiveness of LLM architectures.

2844. Hyperbolic Genome Embeddings

链接：https://iclr.cc/virtual/2025/poster/29862 abstract： Current approaches to genomic sequence modeling often struggle to align the inductive biases of machine learning models with the evolutionarily-informed structure of biological systems. To this end, we formulate a novel application of hyperbolic CNNs that exploits this structure, enabling more expressive DNA sequence representations. Our strategy circumvents the need for explicit phylogenetic mapping while discerning key properties of sequences pertaining to core functional and regulatory behavior. Across 37 out of 42 genome interpretation benchmark datasets, our hyperbolic models outperform their Euclidean equivalents. Notably, our approach even surpasses state-of-the-art performance on seven GUE benchmark datasets, consistently outperforming many DNA language models while using orders of magnitude fewer parameters and avoiding pretraining. Our results include a novel set of benchmark datasets---the Transposable Elements Benchmark---which explores a major but understudied component of the genome with deep evolutionary significance. We further motivate our work by exploring how our hyperbolic models recognize genomic signal under various data-generating conditions and by constructing an empirical method for interpreting the hyperbolicity of dataset embeddings. Throughout these assessments, we find persistent evidence highlighting the potential of our hyperbolic framework as a robust paradigm for genome representation learning. Our code and benchmark datasets are available at https://github.com/rrkhan/HGE.

2845. Triples as the Key: Structuring Makes Decomposition and Verification Easier in LLM-based TableQA

链接：https://iclr.cc/virtual/2025/poster/29439 abstract： As the mainstream approach, LLMs have been widely applied and

researched in TableQA tasks. Currently, the core of LLM-based TableQA methods typically include three phases: question decomposition, sub-question TableQA reasoning, and answer verification. However, several challenges remain in this process: i) Sub-questions generated by these methods often exhibit significant gaps with the original question due to critical information overlooked during the LLM's direct decomposition; ii) Verification of answers is typically challenging because LLMs tend to generate optimal responses during self-correct. To address these challenges, we propose a Triple-Inspired Decomposition and vErification (TIDE) strategy, which leverages the structural properties of triples to assist in decomposition and verification in TableQA. The inherent structure of triples (head entity, relation, tail entity) requires the LLM to extract as many entities and relations from the question as possible. Unlike direct decomposition methods that may overlook key information, our transformed sub-questions using triples encompass more critical details. Additionally, this explicit structure facilitates verification. By comparing the triples derived from the answers with those from the question decomposition, we can achieve easier and more straightforward validation than when relying on the LLM's self-correct tendencies. By employing triples alongside established LLM modes, Direct Prompting and Agent modes, TIDE achieves state-of-the-art performance across multiple TableQA datasets, demonstrating the effectiveness of our method.

2846. Data Pruning by Information Maximization

链接：https://iclr.cc/virtual/2025/poster/30725 abstract： In this paper, we present InfoMax, a novel data pruning method, also known as coreset selection, designed to maximize the information content of selected samples while minimizing redundancy. By doing so, InfoMax enhances the overall informativeness of the coreset. The information of individual samples is measured by importance scores, which capture their influence or difficulty in model learning. To quantify redundancy, we use pairwise sample similarities, based on the premise that similar samples contribute similarly to the learning process.We formalize the coreset selection problem as a discrete quadratic programming (DQP) task, with the objective of maximizing the total information content, represented as the sum of individual sample contributions minus the redundancies introduced by similar samples within the coreset.To ensure practical scalability, we introduce an efficient gradient-based solver, complemented by sparsification techniques applied to the similarity matrix and dataset partitioning strategies. This enables InfoMax to seamlessly scale to datasets with millions of samples. Extensive experiments demonstrate the superior performance of InfoMax in various data pruning tasks, including image classification, vision-language pre-training, and instruction tuning for large language models.

2847. An Evolved Universal Transformer Memory

链接：https://iclr.cc/virtual/2025/poster/28152 abstract： Prior methods propose to offset the escalating costs of modern foundation models by dropping specific parts of their contexts with hand-designed rules, while attempting to preserve their original performance. We overcome this trade-off with Neural Attention Memory Models (NAMMs), introducing a learned network for memory management that improves both the performance and efficiency of transformers. We evolve NAMMs atop pre-trained transformers to provide different latent contexts focusing on the most relevant information for individual layers and attention heads. NAMMs are universally applicable to any model using self-attention as they condition exclusively on the values in the produced attention matrices. Learning NAMMs on a small set of problems, we achieve substantial performance improvements across multiple long-context benchmarks while cutting the model's input contexts up to a fraction of the original sizes. We show the generality of our conditioning enables zero-shot transfer of NAMMs trained only on language to entirely new transformer architectures even across input modalities, with their benefits carrying over to vision and reinforcement learning.

2848. VLAS: Vision-Language-Action Model with Speech Instructions for Customized Robot Manipulation

链接：https://iclr.cc/virtual/2025/poster/30076 abstract： Vision-language-action models (VLAs) have recently become highly prevalent in robot manipulation due to its end-to-end architecture and impressive performance. However, current VLAs are limited to processing human instructions in textual form, neglecting the more natural speech modality for human interaction. A typical approach of incorporating speech modality into VLA necessitates a separate speech recognition system to transcribe spoken instructions into text. Such a cascading pipeline raises two major concerns for robotic systems. First, the entire model grows in size and complexity, potentially resulting in redundant computations and increased memory consumption. Second, the transcription procedure would lose non-semantic information in the raw speech, such as voiceprint, which is crucial for a robot to successfully understand and complete customized tasks. To this end, we propose VLAS, the fisrt end-to-end policy model that seamlessly integrates speech modality for robot manipulation. We present a three-stage speech instruction tuning strategy leveraging multimodal datasets, including our manually curated SQA and CSI datasets. Furthermore, to facilitate personalized operations, we develop a voice retrieval-augmented generation (RAG) approach to enhance the robot's performance in tasks requiring individual-specific knowledge. Experimental results show that the proposed VLAS, following either textual or speech instructions, can achieve performance comparable to traditional VLAs on the CALVIN benchmark. In addition, we created a benchmark consisting of customization tasks, where our VLAS demonstrates absolute superiority by fully leveraging the auxiliary information in speech.

2849. Memory Mosaics

链接：https://iclr.cc/virtual/2025/poster/30157 abstract： Memory Mosaics are networks of associative memories working in concert to achieve a prediction task of interest. Like transformers, memory mosaics possess compositional capabilities and in-context learning capabilities. Unlike transformers, memory mosaics achieve these capabilities in comparatively transparent way (“predictive disentanglement”). We illustrate these capabilities on a toy example and also show that memory mosaics perform as

well or better than transformers on medium-scale language modeling tasks.

2850. X-NeMo: Expressive Neural Motion Reenactment via Disentangled Latent Attention

链接：https://iclr.cc/virtual/2025/poster/29937 abstract： We propose X-NeMo, a novel zero-shot diffusion-based portrait animation pipeline that animates a static portrait using facial movements from a driving video of a different individual. Our work first identifies the root causes of the limitations in prior approaches, such as identity leakage and difficulty in capturing subtle and extreme expressions. To address these challenges, we introduce a fully end-to-end training framework that distills a 1D identity-agnostic latent motion descriptor from driving image, effectively controlling motion through cross-attention during image generation. Our implicit motion descriptor captures expressive facial motion in fine detail, learned end-to-end from a diverse video dataset without reliance on any pre-trained motion detectors. We further disentangle motion latents from identity cues with enhanced expressiveness by supervising their learning with a dual GAN decoder, alongside spatial and color augmentations. By embedding the driving motion into a 1D latent vector and controlling motion via cross-attention instead of additive spatial guidance, our design effectively eliminates the transmission of spatial-aligned structural clues from the driving condition to the diffusion backbone, substantially mitigating identity leakage. Extensive experiments demonstrate that X-NeMo surpasses state-of-the-art baselines, producing highly expressive animations with superior identity resemblance. Our code and models will be available for research.

2851. Gradient-Free Generation for Hard-Constrained Systems

链接：https://iclr.cc/virtual/2025/poster/28037 abstract： Generative models that satisfy hard constraints are critical in many scientific and engineering applications, where physical laws or system requirements must be strictly respected. Many existing constrained generative models, especially those developed for computer vision, rely heavily on gradient information, which is often sparse or computationally expensive in some fields, e.g., partial differential equations (PDEs). In this work, we introduce a novel framework for adapting pre-trained, unconstrained flow-matching models to satisfy constraints exactly in a zero-shot manner without requiring expensive gradient computations or fine-tuning. Our framework, ECI sampling, alternates between extrapolation (E), correction (C), and interpolation (I) stages during each iterative sampling step of flow matching sampling to ensure accurate integration of constraint information while preserving the validity of the generation. We demonstrate the effectiveness of our approach across various PDE systems, showing that ECI-guided generation strictly adheres to physical constraints and accurately captures complex distribution shifts induced by these constraints. Empirical results demonstrate that our framework consistently outperforms baseline approaches in various zero-shot constrained generation tasks and also achieves competitive results in the regression tasks without additional fine-tuning.

2852. CityGaussianV2: Efficient and Geometrically Accurate Reconstruction for Large-Scale Scenes

链接：https://iclr.cc/virtual/2025/poster/29187 abstract： Recently, 3D Gaussian Splatting (3DGS) has revolutionized radiance field reconstruction, manifesting efficient and high-fidelity novel view synthesis. However, accurately representing surfaces, especially in large and complex scenarios, remains a significant challenge due to the unstructured nature of 3DGS. In this paper, we present CityGaussianV2, a novel approach for large-scale scene reconstruction that addresses critical challenges related to geometric accuracy and efficiency. Building on the favorable generalization capabilities of 2D Gaussian Splatting (2DGS), we address its convergence and scalability issues. Specifically, we implement a decomposed-gradient-based densification and depth regression technique to eliminate blurry artifacts and accelerate convergence. To scale up, we introduce an elongation filter that mitigates Gaussian count explosion caused by 2DGS degeneration. Furthermore, we optimize the CityGaussian pipeline for parallel training, achieving up to 10$\times$ compression, at least 25\% savings in training time, and a 50\% decrease in memory usage. We also established standard geometry benchmarks under large-scale scenes. Experimental results demonstrate that our method strikes a promising balance between visual quality, geometric accuracy, as well as storage and training costs.

2853. FreeVS: Generative View Synthesis on Free Driving Trajectory

链接：https://iclr.cc/virtual/2025/poster/28983 abstract： Existing reconstruction-based novel view synthesis methods for driving scenes focus on synthesizing camera views along the recorded trajectory of the ego vehicle. Their image rendering performance will severely degrade on viewpoints falling out of the recorded trajectory, where camera rays are untrained.We propose FreeVS, a novel fully generative approach that can synthesize camera views on free new trajectories in real driving scenes. To control the generation results to be 3D consistent with the real scenes and accurate in viewpoint pose, we propose the pseudo-image representation of view priors to control the generation process.Viewpoint translation simulation is applied on pseudo-images to simulate camera movement in each direction.Once trained, FreeVS can be applied to any validation sequences without reconstruction process and synthesis views on novel trajectories.Moreover, we propose two new challenging benchmarks tailored to driving scenes, which are novel camera synthesis and novel trajectory synthesis, emphasizing the freedom of viewpoints.Given that no ground truth images are available on novel trajectories, we also propose to evaluate the consistency of images synthesized on novel trajectories with 3D perception models.Experiments on the Waymo Open Dataset show that FreeVS has a strong image synthesis performance on both the recorded trajectories and novel trajectories. The code is released. Project page: https://freevs24.github.io/.

2854. Improving Long-Text Alignment for Text-to-Image Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/31132 abstract： The rapid advancement of text-to-image (T2I) diffusion models has enabled them to generate unprecedented results from given texts. However, as text inputs become longer, existing encoding methods like CLIP face limitations, and aligning the generated images with long texts becomes challenging. To tackle these issues, we propose LongAlign, which includes a segment-level encoding method for processing long texts and a decomposed preference optimization method for effective alignment training. For segment-level encoding, long texts are divided into multiple segments and processed separately. This method overcomes the maximum input length limits of pretrained encoding models. For preference optimization, we provide decomposed CLIP-based preference models to fine-tune diffusion models. Specifically, to utilize CLIP-based preference models for T2I alignment, we delve into their scoring mechanisms and find that the preference scores can be decomposed into two components: a text-relevant part that measures T2I alignment and a text-irrelevant part that assesses other visual aspects of human preference. Additionally, we find that the text-irrelevant part contributes to a common overfitting problem during fine-tuning. To address this, we propose a reweighting strategy that assigns different weights to these two components, thereby reducing overfitting and enhancing alignment. After fine-tuning 512×512 Stable Diffusion (SD) v1.5 for about 20 hours using our method, the fine-tuned SD outperforms stronger foundation models in T2I alignment, such as PixArt-α and Kandinsky v2.2. The code is available at https://github.com/luping-liu/LongAlign.

2855. Enhancing End-to-End Autonomous Driving with Latent World Model

链接：https://iclr.cc/virtual/2025/poster/28868 abstract： In autonomous driving, end-to-end planners directly utilize raw sensor data, enabling them to extract richer scene features and reduce information loss compared to traditional planners. This raises a crucial research question: how can we develop better scene feature representations to fully leverage sensor data in end-to-end driving? Self-supervised learning methods show great success in learning rich feature representations in NLP and computer vision. Inspired by this, we propose a novel self-supervised learning approach using the LAtent World model (LAW) for end-to-end driving. LAW predicts future latent scene features based on current features and ego trajectories. This self-supervised task can be seamlessly integrated into perception-free and perception-based frameworks, improving scene feature learning while optimizing trajectory prediction. LAW achieves state-of-the-art performance across multiple benchmarks, including real-world open-loop benchmark nuScenes, NAVSIM, and simulator-based closed-loop benchmark CARLA. The code will be released.

2856. On the Computation of the Fisher Information in Continual Learning

链接：https://iclr.cc/virtual/2025/poster/31332 abstract： One of the most popular methods for continual learning with deep neural networks is Elastic Weight Consolidation (EWC), which involves computing the Fisher Information. The exact way in which the Fisher Information is computed is however rarely described, and multiple different implementations for it can be found online. This blog post discusses and empirically compares several often-used implementations, which highlights that many currently reported results for EWC could likely be improved by changing the way the Fisher Information is computed.

2857. DeFT: Decoding with Flash Tree-attention for Efficient Tree-structured LLM Inference

链接：https://iclr.cc/virtual/2025/poster/31131 abstract： Large language models (LLMs) are increasingly employed for complex tasks that process multiple generation calls in a tree structure with shared prefixes of tokens, including few-shot prompting, multi-step reasoning, speculative decoding, etc. However, existing inference systems for tree-based applications are inefficient due to improper partitioning of queries and KV cache during attention calculation.This leads to two main issues: (1) a lack of memory access (IO) reuse for KV cache of shared prefixes, and (2) poor load balancing.As a result, there is redundant KV cache IO between GPU global memory and shared memory, along with low GPU utilization. To address these challenges, we propose DeFT(Decoding with Flash Tree-Attention), a hardware-efficient attention algorithm with prefix-aware and load-balanced KV cache partitions. DeFT reduces the number of read/write operations of KV cache during attention calculation through **KV-Guided Grouping**, a method that avoids repeatedly loading KV cache of shared prefixes in attention computation. Additionally, we propose **Flattened Tree KV Splitting**, a mechanism that ensures even distribution of the KV cache across partitions with little computation redundancy, enhancing GPU utilization during attention computations. By reducing 73-99% KV cache IO and nearly 100% IO for partial results during attention calculation, DeFT achieves up to 2.23/3.59$\times$ speedup in the end-to-end/attention latency across three practical tree-based workloads compared to state-of-the-art attention algorithms. Our code is available at https://github.com/LINs-lab/DeFT.

2858. SimulPL: Aligning Human Preferences in Simultaneous Machine Translation

链接：https://iclr.cc/virtual/2025/poster/29316 abstract： Simultaneous Machine Translation (SiMT) generates translations while receiving streaming source inputs. This requires the SiMT model to learn a read/write policy, deciding when to translate and when to wait for more source input. Numerous linguistic studies indicate that audiences in SiMT scenarios have distinct preferences, such as accurate translations, simpler syntax, and no unnecessary latency. Aligning SiMT models with these human preferences is crucial to improve their performances. However, this issue still remains unexplored. Additionally, preference optimization for SiMT task is also challenging. Existing methods focus solely on optimizing the generated responses, ignoring

human preferences related to latency and the optimization of read/write policy during the preference optimization phase. To address these challenges, we propose Simultaneous Preference Learning (SimulPL), a preference learning framework tailored for the SiMT task. In the SimulPL framework, we categorize SiMT human preferences into five aspects: **translation quality preference**, **monotonicity preference**, **key point preference**, **simplicity preference**, and **latency preference**. By leveraging the first four preferences, we construct human preference prompts to efficiently guide GPT-4/4o in generating preference data for the SiMT task. In the preference optimization phase, SimulPL integrates **latency preference** into the optimization objective and enables SiMT models to improve the read/write policy, thereby aligning with human preferences more effectively. Experimental results indicate that SimulPL exhibits better alignment with human preferences across all latency levels in Zh$\rightarrow$En, De$\rightarrow$En and En$\rightarrow$Zh SiMT tasks. Our data and code will be available at https://github.com/EurekaForNLP/SimulPL.

2859. CREAM: Consistency Regularized Self-Rewarding Language Models

链接：https://iclr.cc/virtual/2025/poster/29403 abstract： Recent self-rewarding large language models (LLM) have successfully applied LLM-as-a-Judge to iteratively improve the alignment performance without the need of human annotations for preference data. These methods commonly utilize the same LLM to act as both the policy model (which generates responses) and the reward model (which scores and ranks those responses). The ranked responses are then used as preference pairs to train the LLM via direct alignment technologies (e.g. DPO). However, it is noteworthy that throughout this process, there is no guarantee of accuracy in the rewarding and ranking, which is critical for ensuring accurate rewards and high-quality preference data. Empirical results from relatively small LLMs (e.g., 7B parameters) also indicate that improvements from self-rewarding may diminish after several iterations in certain situations, which we hypothesize is due to accumulated bias in the reward system. This bias can lead to unreliable preference data for training the LLM. To address this issue, we first formulate and analyze the generalized iterative preference fine-tuning framework for self-rewarding language model. We then introduce the regularization to this generalized framework to mitigate the overconfident preference labeling in the self-rewarding process. Based on this theoretical insight, we propose a Consistency Regularized sElf-rewarding lAnguage Model (CREAM) that leverages the consistency of rewards across different iterations to regularize the self-rewarding training, helping the model to learn from more reliable preference data. With this explicit regularization, our empirical results demonstrate the superiority of CREAM in improving both reward consistency and alignment performance. The code is publicly available at https://github.com/Raibows/CREAM.

2860. Beware of Calibration Data for Pruning Large Language Models

链接：https://iclr.cc/virtual/2025/poster/27801 abstract： As large language models (LLMs) are widely applied across various fields, modelcompression has become increasingly crucial for reducing costs and improvinginference efficiency. Post-training pruning is a promising method that does notrequire resource-intensive iterative training and only needs a small amount ofcalibration data to assess the importance of parameters. Recent research has enhanced post-training pruning from different aspects but few of them systematicallyexplore the effects of calibration data, and it is unclear if there exist better calibration data construction strategies. We fill this blank and surprisingly observe thatcalibration data is also crucial to post-training pruning, especially for high sparsity. Through controlled experiments on important influence factors of calibrationdata, including the pruning settings, the amount of data, and its similarity withpre-training data, we observe that a small size of data is adequate, and more similar data to its pre-training stage can yield better performance. As pre-training datais usually inaccessible for advanced LLMs, we further provide a self-generatingcalibration data synthesis strategy to construct feasible calibration data. Experimental results on recent strong open-source LLMs (e.g., DCLM, and LLaMA-3)show that the proposed strategy can enhance the performance of strong pruningmethods (e.g., Wanda, DSnoT, OWL) by a large margin (up to 2.68%).

2861. Rectified Diffusion: Straightness Is Not Your Need in Rectified Flow

链接：https://iclr.cc/virtual/2025/poster/28415 abstract： Diffusion models have greatly improved visual generation but are hindered by slow generation speed due to the computationally intensive nature of solving generative ODEs. Rectified flow, a widely recognized solution, improves generation speed by straightening the ODE path. Its key components include: 1) using the linear interpolating diffusion form of flow-matching, 2) employing $\boldsymbol v$-prediction, and 3) performing rectification (a.k.a. reflow). In this paper, we argue that the success of rectification primarily lies in using a pretrained diffusion model to obtain matched pairs of noise and samples, followed by retraining with these matched noise-sample pairs. Based on this, components 1) and 2) are unnecessary. Furthermore, we highlight that straightness is not an essential training target for rectification; rather, it is a specific case of flow-matching models. The more critical training target is to achieve a first-order approximate ODE path, which is inherently curved for models like DDPM and Sub-VP. Building on this insight, we propose Rectified Diffusion, which generalizes the design space and application scope of rectification to encompass the broader category of diffusion models, rather than being restricted to flow-matching models. We validate our method on Stable Diffusion v1-5 and Stable Diffusion XL. Our method not only greatly simplifies the training procedure of rectified flow-based previous works (e.g., InstaFlow) but also achieves superior performance with even lower training cost. Our code is available at https://github.com/G-U-N/Rectified-Diffusion.

2862. 3DMolFormer: A Dual-channel Framework for Structure-based Drug Discovery

链接：https://iclr.cc/virtual/2025/poster/29642 abstract： Structure-based drug discovery, encompassing the tasks of protein-ligand docking and pocket-aware 3D drug design, represents a core challenge in drug discovery. However, no existing work can deal with both tasks to effectively leverage the duality between them, and current methods for each task are hindered by challenges in modeling 3D information and the limitations of available data. To address these issues, we propose 3DMolFormer, a unified dual-channel transformer-based framework applicable to both docking and 3D drug design tasks, which exploits their duality by utilizing docking functionalities within the drug design process. Specifically, we represent 3D pocket-ligand complexes using parallel sequences of discrete tokens and continuous numbers, and we design a corresponding dual-channel transformer model to handle this format, thereby overcoming the challenges of 3D information modeling. Additionally, we alleviate data limitations through large-scale pre-training on a mixed dataset, followed by supervised and reinforcement learning fine-tuning techniques respectively tailored for the two tasks. Experimental results demonstrate that 3DMolFormer outperforms previous approaches in both protein-ligand docking and pocket-aware 3D drug design, highlighting its promising application in structure-based drug discovery. The code is available at: https://github.com/HXYfighter/3DMolFormer.

2863. Does Editing Provide Evidence for Localization?

链接：https://iclr.cc/virtual/2025/poster/31325 abstract： A basic aspiration for interpretability research in large language models is to localize semantically meaningful behaviors to particular components within the LLM. There are various heuristics for finding candidate locations within the LLM. Once a candidate localization is found, it can be assessed by editing the internal representations at the corresponding localization and checking whether this induces model behavior that is consistent with the semantic interpretion of the localization. The question we address here is, how strong is the evidence provided by such edits? To assess localization, we want to assess the effect of the optimal intervention at a particular location. The key new technical tool is a way of adapting LLM alignment techniques to find such optimal localized edits. With this tool in hand, we give an example where the edit-based evidence for localization appears strong, but where localization clearly fails. Indeed, we find that optimal edits at random localizations can be as effective as aligning the full model. In aggregate, our results suggest that merely observing that localized edits induce targeted changes in behavior provides little to no evidence that these locations actually encode the target behavior.

2864. A Geometric Framework for Understanding Memorization in Generative Models

链接：https://iclr.cc/virtual/2025/poster/29158 abstract： As deep generative models have progressed, recent work has shown them to be capable of memorizing and reproducing training datapoints when deployed. These findings call into question the usability of generative models, especially in light of the legal and privacy risks brought about by memorization. To better understand this phenomenon, we propose the *manifold memorization hypothesis* (MMH), a geometric framework which leverages the manifold hypothesis into a clear language in which to reason about memorization. We propose to analyze memorization in terms of the relationship between the dimensionalities of (i) the ground truth data manifold and (ii) the manifold learned by the model. This framework provides a formal standard for "how memorized" a datapoint is and systematically categorizes memorized data into two types: memorization driven by overfitting and memorization driven by the underlying data distribution. By analyzing prior work in the context of the MMH, we explain and unify assorted observations in the literature. We empirically validate the MMH using synthetic data and image datasets up to the scale of Stable Diffusion, developing new tools for detecting and preventing generation of memorized samples in the process.

2865. TLDR: Token-Level Detective Reward Model for Large Vision Language Models

链接：https://iclr.cc/virtual/2025/poster/29193 abstract： Although reward models have been successful in improving multimodal large language models, the reward models themselves remain brutal and contain minimal information. Notably, existing reward models only mimic human annotations by assigning only one feedback to any text, no matter how long the text is. In the realm of multimodal language models, where models are required to process both images and texts, a naive reward model may learn implicit biases toward texts and become less grounded in images. In this paper, we propose a Token-Level Detective Reward Model (TLDR) to provide fine-grained annotations to each text token. We first introduce a perturbation-based method to generate synthetic hard negatives and their token-level labels to train TLDR models. Then we show the rich usefulness of TLDR models both in assisting off-the-shelf models to self-correct their generations, and in serving as a hallucination evaluation tool. We show that TLDR automatically trains a token-level likelihood optimization, and can improve the base model's performance significantly. Finally, we show that TLDR models can significantly speed up human annotation by 3 times to acquire a broader range of high-quality vision language data.

2866. Glimpse: Enabling White-Box Methods to Use Proprietary Models for Zero-Shot LLM-Generated Text Detection

链接：https://iclr.cc/virtual/2025/poster/29149 abstract： Advanced large language models (LLMs) can generate text almost indistinguishable from human-written text, highlighting the importance of LLM-generated text detection. However, current zero-shot techniques face challenges as white-box methods are restricted to use weaker open-source LLMs, and black-box methods are limited by partial observation from stronger proprietary LLMs. It seems impossible to enable white-box methods to use

proprietary models because API-level access to the models neither provides full predictive distributions nor inner embeddings. To traverse the divide, we propose Glimpse, a probability distribution estimation approach, predicting the full distributions from partial observations. Despite the simplicity of Glimpse, we successfully extend white-box methods like Entropy, Rank, Log-Rank, and Fast-DetectGPT to latest proprietary models. Experiments show that Glimpse with Fast-DetectGPT and GPT-3.5 achieves an average AUROC of about 0.95 in five latest source models, improving the score by 51\% relative to the remaining space of the open source baseline. It demonstrates that the latest LLMs can effectively detect their own outputs, suggesting that advanced LLMs may be the best shield against themselves. We release our code and data at https://github.com/baoguangsheng/glimpse.

2867. SqueezeAttention: 2D Management of KV-Cache in LLM Inference via Layer-wise Optimal Budget

链接：https://iclr.cc/virtual/2025/poster/30707 abstract： Optimizing the Key-Value (KV) cache of the Large Language Model (LLM) has been considered critical to saving the cost of inference. Most of the existing KV-cache compression algorithms attempted to sparsify the sequence of tokens by taking advantage of the different importance of tokens. However, most of these methods treat all layers equally, allocating the same KV budget to each layer. This approach is suboptimal, as some layers may be less sensitive to input tokens yet still receive the same budget as others. In this work, we found that by identifying the importance of attention layers, we could optimize the KV-cache jointly from two dimensions, i.e., sequence-wise and layer-wise. Based on our observations regarding layer-wise importance in inference, we propose \sys to precisely optimize the allocation of KV-cache budget among layers on-the-fly and then incorporate three representative sequence-wise algorithms to compress the KV-cache for each layer with its very own budget. Specifically, we first measure each layer's importance by calculating the cosine similarity of the input prompt differences before and after the self-attention layers. Based on this similarity, we then categorize the layers into two groups and adjust their KV budgets accordingly. By optimizing the KV-cache from both sequence's and layer's dimensions, \sys achieves around 30\% to 70\% of the memory reductions and up to 2.2 $\times$ of throughput improvements in a wide range of LLMs and benchmarks. The code is available at https://github.com/hetailang/SqueezeAttention.

2868. Understanding and Enhancing Safety Mechanisms of LLMs via Safety-Specific Neuron

链接：https://iclr.cc/virtual/2025/poster/27728 abstract： Safety alignment for large language models (LLMs) has become a critical issue due to their rapid progress. However, our understanding of effective safety mechanisms in LLMs remains limited, leading to safety alignment training that mainly focuses on improving optimization, data-level enhancement, or adding extra structures to intentionally block harmful outputs. To address this gap, we develop a neuron detection method to identify safety neurons—those consistently crucial for handling and defending against harmful queries. Our findings reveal that these safety neurons constitute less than 1% of all parameters, are language-specific and are predominantly located in self-attention layers. Moreover, safety is collectively managed by these neurons in the first several layers. Based on these observations, we introduce a $\underline{S}$afety $\underline{N}$euron $\underline{Tun}$ing method, named $\texttt{SN-Tune}$, that exclusively tune safety neurons without compromising models' general capabilities. $\texttt{SN-Tune}$ significantly enhances the safety of instruction-tuned models, notably reducing the harmful scores of Llama3-8B-Instruction from 65.5 to 2.0, Mistral-7B-Instruct-v0.2 from 70.8 to 4.5, and Vicuna-13B-1.5 from 93.5 to 3.0. Moreover, $\texttt{SN-Tune}$ can be applied to base models on efficiently establishing LLMs' safety mechanism. In addition, we propose $\underline{R}$obust $\underline{S}$afety $\underline{N}$euron $\underline{Tun}$ing method ($\texttt{RSN-Tune}$), which preserves the integrity of LLMs' safety mechanisms during downstream task fine-tuning by separating the safety neurons from models' foundation neurons.

2869. ReMoE: Fully Differentiable Mixture-of-Experts with ReLU Routing

链接：https://iclr.cc/virtual/2025/poster/31028 abstract：

2870. Towards Domain Adaptive Neural Contextual Bandits

链接：https://iclr.cc/virtual/2025/poster/30002 abstract： Contextual bandit algorithms are essential for solving real-world decision making problems. In practice, collecting a contextual bandit's feedback from different domains may involve different costs. For example, measuring drug reaction from mice (as a source domain) and humans (as a target domain). Unfortunately, adapting a contextual bandit algorithm from a source domain to a target domain with distribution shift still remains a major challenge and largely unexplored. In this paper, we introduce the first general domain adaptation method for contextual bandits. Our approach learns a bandit model for the target domain by collecting feedback from the source domain. Our theoretical analysis shows that our algorithm maintains a sub-linear regret bound even adapting across domains. Empirical results show that our approach outperforms the state-of-the-art contextual bandit algorithms on real-world datasets. Code will soon be available at https://github.com/Wang-ML-Lab/DABand.

2871. Investigating Pattern Neurons in Urban Time Series Forecasting

链接：https://iclr.cc/virtual/2025/poster/29185 abstract： Urban time series forecasting is crucial for smart city development and is key to sustainable urban management. Although urban time series models (UTSMs) are effective in general forecasting, they

often overlook low-frequency events, such as holidays and extreme weather, leading to degraded performance in practical applications. In this paper, we first investigate how UTSMs handle these infrequent patterns from a neural perspective. Based on our findings, we propose $\textbf{P}$attern $\textbf{N}$euron guided $\textbf{Train}$ing ($\texttt{PN-Train}$), a novel training method that features (i) a $\textit{perturbation-based detector}$ to identify neurons responsible for low-frequency patterns in UTSMs, and (ii) a $\textit{fine-tuning mechanism}$ that enhances these neurons without compromising representation learning on high-frequency patterns. Empirical results demonstrate that $\texttt{PN-Train}$ considerably improves forecasting accuracy for low-frequency events while maintaining high performance for high-frequency events. The code is available at https://github.com/cwang-nus/PN-Train.

2872. Transformers Learn to Implement Multi-step Gradient Descent with Chain of Thought

链接：https://iclr.cc/virtual/2025/poster/28223 abstract： Chain of Thought (CoT) prompting has been shown to significantly improve the performance of large language models (LLMs), particularly in arithmetic and reasoning tasks, by instructing the model to produce intermediate reasoning steps. Despite the remarkable empirical success of CoT and its theoretical advantages in enhancing expressivity, the mechanisms underlying CoT training remain largely unexplored. In this paper, we study the training dynamics of transformers over a CoT objective on a in-context weight prediction task for linear regression. We prove that while a one-layer linear transformer without CoT can only implement a single step of gradient descent (GD) and fails to recover the ground-truth weight vector, a transformer with CoT prompting can learn to perform multi-step GD autoregressively, achieving near-exact recovery. Furthermore, we show that the trained transformer effectively generalizes on the unseen data. Empirically, we demonstrate that CoT prompting yields substantial performance improvements.

2873. DARE the Extreme: Revisiting Delta-Parameter Pruning For Fine-Tuned Models

链接：https://iclr.cc/virtual/2025/poster/29137 abstract： Storing open-source fine-tuned models separately introduces redundancy and increases response times in applications utilizing multiple models. Delta-parameter pruning (DPP), particularly the random drop and rescale (DARE) method proposed by Yu et al., addresses this by pruning the majority of delta parameters—the differences between fine-tuned and pre-trained model weights—while typically maintaining minimal performance loss. However, DARE fails when either the pruning rate or the magnitude of the delta parameters is large. We highlight two key reasons for this failure: (1) an excessively large rescaling factor as pruning rates increase, and (2) high mean and variance in the delta parameters. To push DARE's limits, we introduce DAREx (DARE the eXtreme), which features two algorithmic improvements: (1) DAREx-q, a rescaling factor modification that significantly boosts performance at high pruning rates (e.g., > 30% on COLA and SST2 for encoder models, with even greater gains in decoder models), and (2) DAREx-L2, which combines DARE with AdamR, an in-training method that applies appropriate delta regularization before DPP. We also demonstrate that DAREx-q can be seamlessly combined with vanilla parameter-efficient fine-tuning techniques like LoRA and can facilitate structural DPP. Additionally, we revisit the application of importance-based pruning techniques within DPP, demonstrating that they outperform random-based methods when delta parameters are large. Through this comprehensive study, we develop a pipeline for selecting the most appropriate DPP method under various practical scenarios.

2874. Adding Conditional Control to Diffusion Models with Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28089 abstract： Diffusion models are powerful generative models that allow for precise control over the characteristics of the generated samples. While these diffusion models trained on large datasets have achieved success, there is often a need to introduce additional controls in downstream fine-tuning processes, treating these powerful models as pre-trained diffusion models. This work presents a novel method based on reinforcement learning (RL) to add such controls using an offline dataset comprising inputs and labels. We formulate this task as an RL problem, with the classifier learned from the offline dataset and the KL divergence against pre-trained models serving as the reward functions. Our method, CTRL (Conditioning pre-Trained diffusion models with Reinforcement Learning), produces soft-optimal policies that maximize the abovementioned reward functions. We formally demonstrate that our method enables sampling from the conditional distribution with additional controls during inference.Our RL-based approach offers several advantages over existing methods. Compared to classifier-free guidance,it improves sample efficiency and can greatly simplify dataset construction by leveraging conditional independence between the inputs and additional controls. Additionally, unlike classifier guidance, it eliminates the need to train classifiers from intermediate states to additional controls.The code is available at https://github.com/zhaoyl18/CTRL.

2875. Distilling Dataset into Neural Field

链接：https://iclr.cc/virtual/2025/poster/28420 abstract： Utilizing a large-scale dataset is essential for training high-performance deep learning models, but it also comes with substantial computation and storage costs. To overcome these challenges, dataset distillation has emerged as a promising solution by compressing the large-scale dataset into a smaller synthetic dataset that retains the essential information needed for training. This paper proposes a novel parameterization framework for dataset distillation, coined Distilling Dataset into Neural Field (DDiF), which leverages the neural field to store the

necessary information of the large-scale dataset. Due to the unique nature of the neural field, which takes coordinates as input and output quantity, DDiF effectively preserves the information and easily generates various shapes of data. We theoretically confirm that DDiF exhibits greater expressiveness than some previous literature when the utilized budget for a single synthetic instance is the same. Through extensive experiments, we demonstrate that DDiF achieves superior performance on several benchmark datasets, extending beyond the image domain to include video, audio, and 3D voxel. We release the code at \url{https://github.com/aailab-kaist/DDiF}.

2876. Black Sheep in the Herd: Playing with Spuriously Correlated Attributes for Vision-Language Recognition

链接：https://iclr.cc/virtual/2025/poster/28838 abstract： Few-shot adaptation for Vision-Language Models (VLMs) presents a dilemma: balancing in-distribution accuracy with out-of-distribution generalization. Recent research has utilized low-level concepts such as visual attributes to enhance generalization. However, this study reveals that VLMs overly rely on a small subset of attributes on decision-making, which co-occur with the category but are not inherently part of it, termed spuriously correlated attributes. This biased nature of VLMs results in poor generalization. To address this, 1) we first propose Spurious Attribute Probing (SAP), identifying and filtering out these problematic attributes to significantly enhance the generalization of existing attribute-based methods; 2) We introduce Spurious Attribute Shielding (SAS), a plug-and-play module that mitigates the influence of these attributes on prediction, seamlessly integrating into various Parameter-Efficient Fine-Tuning (PEFT) methods. In experiments, SAP and SAS significantly enhance accuracy on distribution shifts across 11 datasets and 3 generalization tasks without compromising downstream performance, establishing a new state-of-the-art benchmark.

2877. Dissecting Adversarial Robustness of Multimodal LM Agents

链接：https://iclr.cc/virtual/2025/poster/29257 abstract： As language models (LMs) are used to build autonomous agents in real environments, ensuring their adversarial robustness becomes a critical challenge. Unlike chatbots, agents are compound systems with multiple components taking actions, which existing LMs safety evaluations do not adequately address. To bridge this gap, we manually create 200 targeted adversarial tasks and evaluation scripts in a realistic threat model on top of VisualWebArena, a real environment for web agents. To systematically examine the robustness of agents, we propose the Agent Robustness Evaluation (ARE) framework. ARE views the agent as a graph showing the flow of intermediate outputs between components and decomposes robustness as the flow of adversarial information on the graph. We find that we can successfully break latest agents that use black-box frontier LMs, including those that perform reflection and tree search. With imperceptible perturbations to a single image (less than 5% of total web page pixels), an attacker can hijack these agents to execute targeted adversarial goals with success rates up to 67%. We also use ARE to rigorously evaluate how the robustness changes as new components are added. We find that inference-time compute that typically improves benign performance can open up new vulnerabilities and harm robustness. An attacker can compromise the evaluator used by the reflexion agent and the value function of the tree search agent, which increases the attack success relatively by 15% and 20%. Our data and code for attacks, defenses, and evaluation are at https://github.com/ChenWu98/agent-attack

2878. Incorporating Visual Correspondence into Diffusion Model for Virtual Try-On

链接：https://iclr.cc/virtual/2025/poster/29302 abstract：

2879. Wayward Concepts In Multimodal Models

链接：https://iclr.cc/virtual/2025/poster/30851 abstract：

2880. SLoPe: Double-Pruned Sparse Plus Lazy Low-Rank Adapter Pretraining of LLMs

链接：https://iclr.cc/virtual/2025/poster/28497 abstract：

2881. Can Watermarks be Used to Detect LLM IP Infringement For Free?

链接：https://iclr.cc/virtual/2025/poster/30061 abstract： The powerful capabilities of LLMs stem from their rich training data and high-quality labeled datasets, making the training of strong LLMs a resource-intensive process, which elevates the importance of IP protection for such LLMs. Compared to gathering high-quality labeled data, directly sampling outputs from these fully trained LLMs as training data presents a more cost-effective approach. This practice—where a suspect model is fine-tuned using high-quality data derived from these LLMs, thereby gaining capabilities similar to the target model—can be seen as a form of IP infringement against the original LLM. In recent years, LLM watermarks have been proposed and used to detect whether a text is AI-generated. Intuitively, if data sampled from a watermarked LLM is used for training, the resulting model would also be influenced by this watermark. This raises the question: can we directly use such watermarks to detect IP infringement of LLMs? In this paper, we explore the potential of LLM watermarks for detecting model infringement. We find that there are two

issues with direct detection: (1) The queries used to sample output from the suspect LLM have a significant impact on detectability. (2) The watermark that is easily learned by LLMs exhibits instability regarding the watermark's hash key during detection. To address these issues, we propose LLDet, a detection method that leverages available anchor LLMs to select suitable queries for sampling from the suspect LLM. Additionally, it adapts the detection threshold to mitigate detection failures caused by different hash keys. To demonstrate the effectiveness of this approach, we construct a challenging model set containing multiple suspect LLMs on which direct detection methods struggle to yield effective results. Our method achieves over 90\% accuracy in distinguishing between infringing and clean models, demonstrating the feasibility of using LLM watermarks to detect LLM IP infringement.

2882. Learning Diagrams: A Graphical Language for Compositional Training Regimes

链接：https://iclr.cc/virtual/2025/poster/28962 abstract： Motivated by deep learning regimes with multiple interacting yet distinct model components, we introduce learning diagrams, graphical depictions of training setups that capture parameterized learning as data rather than code. A learning diagram compiles to a unique loss function on which component models are trained. The result of training on this loss is a collection of models whose predictions ``agree'' with one another. We show that a number of popular learning setups such as few-shot multi-task learning, knowledge distillation, and multi-modal learning can be depicted as learning diagrams. We further implement learning diagrams in a library that allows users to build diagrams of PyTorch and Flux.jl models. By implementing some classic machine learning use cases, we demonstrate how learning diagrams allow practitioners to build complicated models as compositions of smaller components, identify relationships between workflows, and manipulate models during or after training. Leveraging a category theoretic framework, we introduce a rigorous semantics for learning diagrams that puts such operations on a firm mathematical foundation.

2883. EcoFace: Audio-Visual Emotional Co-Disentanglement Speech-Driven 3D Talking Face Generation

链接：https://iclr.cc/virtual/2025/poster/28709 abstract： Speech-driven 3D facial animation has attracted significant attention due to its wide range of applications in animation production and virtual reality. Recent research has explored speech-emotion disentanglement to enhance facial expressions rather than manually assigning emotions. However, this approach face issues such as feature confusion, emotions weakening and mean-face. To address these issues, we present EcoFace, a framework that (1) proposes a novel collaboration objective to provide a explicit signal for emotion representation learning from the speaker's expressive movements and produced sounds, constructing an audio-visual joint and coordinated emotion space that is independent of speech content. (2) constructs a universal facial motion distribution space determined by speech features and implement speaker-specific generation. Extensive experiments show that our method achieves more generalized and emotionally realistic talking face generation compared to previous methods.

2884. Neural Approximate Mirror Maps for Constrained Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/27899 abstract： Diffusion models excel at creating visually-convincing images, but they often struggle to meet subtle constraints inherent in the training data. Such constraints could be physics-based (e.g., satisfying a PDE), geometric (e.g., respecting symmetry), or semantic (e.g., including a particular number of objects). When the training data all satisfy a certain constraint, enforcing this constraint on a diffusion model makes it more reliable for generating valid synthetic data and solving constrained inverse problems. However, existing methods for constrained diffusion models are restricted in the constraints they can handle. For instance, recent work proposed to learn mirror diffusion models (MDMs), but analytical mirror maps only exist for convex constraints and can be challenging to derive. We propose neural approximate mirror maps (NAMMs) for general, possibly non-convex constraints. Our approach only requires a differentiable distance function from the constraint set. We learn an approximate mirror map that transforms data into an unconstrained space and a corresponding approximate inverse that maps data back to the constraint set. A generative model, such as an MDM, can then be trained in the learned mirror space and its samples restored to the constraint set by the inverse map. We validate our approach on a variety of constraints, showing that compared to an unconstrained diffusion model, a NAMM-based MDM substantially improves constraint satisfaction. We also demonstrate how existing diffusion-based inverse-problem solvers can be easily applied in the learned mirror space to solve constrained inverse problems.

2885. Self-MoE: Towards Compositional Large Language Models with Self-Specialized Experts

链接：https://iclr.cc/virtual/2025/poster/30184 abstract： We present Self-MoE, an approach that transforms a monolithic LLM into a compositional, modular system of self-specialized experts, named MiXSE (MiXture of Self-specialized Experts). Our approach leverages self-specialization, which constructs expert modules using self-generated synthetic data, each equipping a shared base LLM with distinct domain-specific capabilities, activated via self-optimized routing. This allows for dynamic and capability-specific handling of various target tasks, enhancing overall capabilities, without extensive human-labeled data and added parameters. Our empirical results reveal that specializing LLMs may exhibit potential trade-offs in performances on non-specialized tasks. On the other hand, our Self-MoE demonstrates substantial improvements (6.5%p on average) over the base LLM across diverse benchmarks such as knowledge, reasoning, math, and coding. It also consistently outperforms other

methods, including instance merging and weight merging, while offering better flexibility and interpretability by design with semantic experts and routing. Our findings highlight the critical role of modularity, the applicability of Self-MoE to multiple base LLMs, and the potential of self-improvement in achieving efficient, scalable, and adaptable systems.

2886. Satisficing Regret Minimization in Bandits

链接：https://iclr.cc/virtual/2025/poster/30941 abstract： Motivated by the concept of satisficing in decision-making, we consider the problem of satisficing exploration in bandit optimization. In this setting, the learner aims at finding a satisficing arm whose mean reward exceeds a certain threshold. The performance is measured by satisficing regret, which is the cumulative deficit of the chosen arm's mean reward compared to the threshold. We propose $\texttt{SELECT}$, a general algorithmic template for Satisficing REgret Minimization via SampLing and LowEr Confidence bound Testing, that attains constant satisficing regret for a wide variety of bandit optimization problems in the realizable case (i.e., whenever a satisficing arm exists). Specifically, given a class of bandit optimization problems and a corresponding learning oracle with sub-linear (standard) regret upper bound, $\texttt{SELECT}$ iteratively makes use of the oracle to identify a potential satisficing arm. Then, it collects data samples from this arm, and continuously compares the lower confidence bound of the identified arm's mean reward against the threshold value to determine if it is a satisficing arm. As a complement, $\texttt{SELECT}$ also enjoys the same (standard) regret guarantee as the oracle in the non-realizable case. Finally, we conduct numerical experiments to validate the performance of $\texttt{SELECT}$ for several popular bandit optimization settings.

2887. Controlling Space and Time with Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29013 abstract： We present 4DiM, a cascaded diffusion model for 4D novel view synthesis (NVS), supporting generation with arbitrary camera trajectories and timestamps, in natural scenes, conditioned on one or more images. With a novel architecture and sampling procedure, we enable training on a mixture of 3D (with camera pose), 4D (pose+time) and video (time but no pose) data, which greatly improves generalization to unseen images and camera pose trajectories over prior works which generally operate in limited domains (e.g., object centric).4DiM is the first-ever NVS method with intuitive metric-scale camera pose control enabled by our novel calibration pipeline for structure-from-motion-posed data. Experiments demonstrate that 4DiM outperforms prior 3D NVS models both in terms of image fidelity and pose alignment, while also enabling the generation of scene dynamics. 4DiM provides a general framework for a variety of tasks including single-image-to-3D, two-image-to-video (interpolation and extrapolation), and pose-conditioned video-to-video translation, which we illustrate qualitatively on a variety of scenes.See https://4d-diffusion.github.io for video samples.

2888. Needle In A Video Haystack: A Scalable Synthetic Evaluator for Video MLLMs

链接：https://iclr.cc/virtual/2025/poster/29224 abstract： Video understanding is a crucial next step for multimodal large language models (MLLMs).Various benchmarks are introduced for better evaluating the MLLMs.Nevertheless, current video benchmarks are still inefficient for evaluating video models during iterative development due to the high cost of constructing datasets and the difficulty in isolating specific skills.In this paper, we propose VideoNIAH (Video Needle in A Haystack), a benchmark construction framework through synthetic video generation. VideoNIAH decouples video content from their query-responses by inserting unrelated visual 'needles' into original videos. The framework automates the generation of query-response pairs using predefined rules, minimizing manual labor. The queries focus on specific aspects of video understanding, enabling more skill-specific evaluations. The separation between video content and the queries also allow for increased video variety and evaluations across different lengths.Utilizing VideoNIAH, we compile a video benchmark, VNBench, which includes tasks such as retrieval, ordering, and counting to evaluate three key aspects of video understanding: temporal perception, chronological ordering, and spatio-temporal coherence. We conduct a comprehensive evaluation of both proprietary and open-source models, uncovering significant differences in their video understanding capabilities across various tasks. Additionally, we perform an in-depth analysis of the test results and model configurations. Based on these findings, we provide some advice for improving video MLLM training, offering valuable insights to guide future research and model development.

2889. Exploring the Design Space of Visual Context Representation in Video MLLMs

链接：https://iclr.cc/virtual/2025/poster/29477 abstract：

2890. GANDALF: Generative AttentioN based Data Augmentation and predictive modeLing Framework for personalized cancer treatment

链接：https://iclr.cc/virtual/2025/poster/29335 abstract： Effective treatment of cancer is a major challenge faced by healthcare providers, due to the highly individualized nature of patient responses to treatment. This is caused by the heterogeneity seen in cancer-causing alterations (mutations) across patient genomes. Limited availability of response data in patients makes it difficult to train personalized treatment recommendation models on mutations from clinical genomic sequencing reports. Prior methods tackle this by utilising larger, labelled pre-clinical laboratory datasets (‘cell lines’), via transfer learning. These methods augment patient data by learning a shared, domain-invariant representation, between the cell line and patient domains, which is

then used to train a downstream drug response prediction (DRP) model. This approach augments data in the shared space but fails to model patient-specific characteristics, which have a strong influence on their drug response. We propose a novel generative attention-based data augmentation and predictive modeling framework, GANDALF, to tackle this crucial shortcoming of prior methods. GANDALF not only augments patient genomic data directly, but also accounts for its domain-specific characteristics. GANDALF outperforms state-of-the-art DRP models on publicly available patient datasets and emerges as the front-runner amongst SOTA cancer DRP models.

2891. Towards Realistic Data Generation for Real-World Super-Resolution

链接：https://iclr.cc/virtual/2025/poster/30091 abstract： Existing image super-resolution (SR) techniques often fail to generalize effectively in complex real-world settings due to the significant divergence between training data and practical scenarios. To address this challenge, previous efforts have either manually simulated intricate physical-based degradations or utilized learning-based techniques, yet these approaches remain inadequate for producing large-scale, realistic, and diverse data simultaneously. In this paper, we introduce a novel Realistic Decoupled Data Generator (RealDGen), an unsupervised learning data generation framework designed for real-world super-resolution. We meticulously develop content and degradation extraction strategies, which are integrated into a novel content-degradation decoupled diffusion model to create realistic low-resolution images from unpaired real LR and HR images. Extensive experiments demonstrate that RealDGen excels in generating large-scale, high-quality paired data that mirrors real-world degradations, significantly advancing the performance of popular SR models on various real-world benchmarks.

2892. Provably Safeguarding a Classifier from OOD and Adversarial Samples

链接：https://iclr.cc/virtual/2025/poster/28554 abstract： This paper aims to transform a trained classifier into an abstaining classifier, suchthat the latter is provably protected from out-of-distribution and adversarial samples. The proposed Sample-efficient Probabilistic Detection using Extreme ValueTheory (SPADE) approach relies on a Generalized Extreme Value (GEV) modelof the training distribution in the latent space of the classifier. Under mild assumptions, this GEV model allows for formally characterizing out-of-distributionand adversarial samples and rejecting them. Empirical validation of the approachis conducted on various neural architectures (ResNet, VGG, and Vision Transformer) and considers medium and large-sized datasets (CIFAR-10, CIFAR-100,and ImageNet). The results show the stability and frugality of the GEV model anddemonstrate SPADE's efficiency compared to the state-of-the-art methods.

2893. Mask-DPO: Generalizable Fine-grained Factuality Alignment of LLMs

链接：https://iclr.cc/virtual/2025/poster/29014 abstract： Large language models (LLMs) exhibit hallucinations (i.e., unfaithful or nonsensical information) when serving as AI assistants in various domains. Since hallucinations always come with truthful content in the LLM responses, previous factuality alignment methods that conduct response-level preference learning inevitably introduced noises during training. Therefore, this paper proposes a fine-grained factuality alignment method based on Direct Preference Optimization (DPO), called Mask-DPO. Incorporating sentence-level factuality as mask signals, Mask-DPO only learns from factually correct sentences in the preferred samples and prevents the penalty on factual contents in the not preferred samples, which resolves the ambiguity in the preference learning. Extensive experimental results demonstrate that Mask-DPO can significantly improve the factuality of LLMs responses to questions from both in-domain and out-of-domain datasets, although these questions and their corresponding topics are unseen during training. Only trained on the ANAH train set, the score of Llama3.1-8B-Instruct on the ANAH test set is improved from 49.19% to 77.53%, even surpassing the score of Llama3.1-70B-Instruct (53.44%), while its FactScore on the out-of-domain Biography dataset is also improved from 30.29% to 39.39%. We further study the generalization property of Mask-DPO using different training sample scaling strategies and find that scaling the number of topics in the dataset is more effective than the number of questions. We provide a hypothesis of what factual alignment is doing with LLMs, on the implication of this phenomenon, and conduct proof-of-concept experiments to verify it. We hope the method and the findings pave the way for future research on scaling factuality alignment.

2894. MindSearch: Mimicking Human Minds Elicits Deep AI Searcher

链接：https://iclr.cc/virtual/2025/poster/27772 abstract： Information seeking and integration is a complex cognitive task that consumes enormous time and effort. Inspired by the remarkable progress of Large Language Models, recent works attempt to solve this task by combining LLMs and search engines. However, these methods still obtain unsatisfying performance due to three challenges: (1) complex requests often cannot be accurately and completely retrieved by the search engine once (2) corresponding information to be integrated is spread over multiple web pages along with massive noise, and (3) a large number of web pages with long contents may quickly exceed the maximum context length of LLMs. Inspired by the cognitive process when humans solve these problems, we introduce MindSearch to mimic the human minds in web information seeking and integration, which can be instantiated by a simple yet effective LLM-based multi-agent framework. The WebPlanner models the human mind of multi-step information seeking as a dynamic graph construction process: it decomposes the user query into atomic sub-questions as nodes in the graph and progressively extends the graph based on the search result from WebSearcher. Tasked with each sub-question, WebSearcher performs hierarchical information retrieval with search engines and collects valuable information for WebPlanner. The multi-agent design of MindSearch enables the whole framework to seek and integrate information parallelly from larger-scale (e.g., more than 300) web pages in 3 minutes, which is worth 3 hours of human effort. MindSearch demonstrates significant improvement in the response quality in terms of depth and breadth, on both close-set and open-set QA problems. Besides, responses from MindSearch based on InternLM2.5-7B are preferable by humans to ChatGPT-

Web and Perplexity.ai applications, which implies that MindSearch can already deliver a competitive solution to the proprietary AI search engine.

2895. Navigation-Guided Sparse Scene Representation for End-to-End Autonomous Driving

链接：https://iclr.cc/virtual/2025/poster/29393 abstract： End-to-End Autonomous Driving (E2EAD) methods typically rely on supervised perception tasks to extract explicit scene information (e.g., objects, maps). This reliance necessitates expensive annotations and constrains deployment and data scalability in real-time applications. In this paper, we introduce SSR, a novel framework that utilizes only 16 navigation-guided tokens as Sparse Scene Representation, efficiently extracting crucial scene information for E2EAD. Our method eliminates the need for human-designed supervised sub-tasks, allowing computational resources to concentrate on essential elements directly related to navigation intent. We further introduce a temporal enhancement module, aligning predicted future scenes with actual future scenes through self-supervision. SSR achieves a 27.2\% relative reduction in L2 error and a 51.6\% decrease in collision rate to UniAD in nuScenes, with a 10.9× faster inference speed and 13× faster training time. Moreover, SSR outperforms VAD-Base with a 48.6-point improvement on driving score in CARLA's Town05 Long benchmark. This framework represents a significant leap in real-time autonomous driving systems and paves the way for future scalable deployment. Code is available at https://github.com/PeidongLi/SSR.

2896. DSBench: How Far Are Data Science Agents from Becoming Data Science Experts?

链接：https://iclr.cc/virtual/2025/poster/30458 abstract： Large Language Models (LLMs) and Large Vision-Language Models (LVLMs) have demonstrated impressive language/vision reasoning abilities, igniting the recent trend of building agents for targeted applications such as shopping assistants or AI software engineers. Recently, many data science benchmarks have been proposed to investigate their performance in the data science domain. However, existing data science benchmarks still fall short when compared to real-world data science applications due to their simplified settings. To bridge this gap, we introduce DSBench, a comprehensive benchmark designed to evaluate data science agents with realistic tasks. This benchmark includes 466 data analysis tasks and 74 data modeling tasks, sourced from Eloquence and Kaggle competitions. DSBench offers a realistic setting by encompassing long contexts, multimodal task backgrounds, reasoning with large data files and multi-table structures, and performing end-to-end data modeling tasks. Our evaluation of state-of-the-art LLMs, LVLMs, and agents shows that they struggle with most tasks, with the best agent solving only 34.12% of data analysis tasks and achieving a 34.74% Relative Performance Gap (RPG). These findings underscore the need for further advancements in developing more practical, intelligent, and autonomous data science agents.

2897. Probe before You Talk: Towards Black-box Defense against Backdoor Unalignment for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30397 abstract： Backdoor unalignment attacks against Large Language Models (LLMs) enable the stealthy compromise of safety alignment using a hidden trigger while evading normal safety auditing. These attacks pose significant threats to the applications of LLMs in the real-world Large Language Model as a Service (LLMaaS) setting, where the deployed model is a fully black-box system that can only interact through text. Furthermore, the sample-dependent nature of the attack target exacerbates the threat. Instead of outputting a fixed label, the backdoored LLM follows the semantics of any malicious command with the hidden trigger, significantly expanding the target space. In this paper, we introduce BEAT, a black-box defense that detects triggered samples during inference to deactivate the backdoor. It is motivated by an intriguing observation (dubbed the probe concatenate effect), where concatenated triggered samples significantly reduce the refusal rate of the backdoored LLM towards a malicious probe, while non-triggered samples have little effect. Specifically, BEAT identifies whether an input is triggered by measuring the degree of distortion in the output distribution of the probe before and after concatenation with the input. Our method addresses the challenges of sample-dependent targets from an opposite perspective. It captures the impact of the trigger on the refusal signal (which is sample-independent) instead of sample-specific successful attack behaviors. It overcomes black-box access limitations by using multiple sampling to approximate the output distribution. Extensive experiments are conducted on various backdoor attacks and LLMs (including the closed-source GPT-3.5-turbo), verifying the effectiveness and efficiency of our defense. Besides, we also preliminarily verify that BEAT can effectively defend against popular jailbreak attacks, as they can be regarded as "natural backdoors". Our source code is available at https://github.com/clearloveclearlove/BEAT.

2898. Uncertainty modeling for fine-tuned implicit functions

链接：https://iclr.cc/virtual/2025/poster/28687 abstract： Implicit functions such as Neural Radiance Fields (NeRFs), occupancy networks, and signed distance functions (SDFs) have become pivotal in computer vision for reconstructing detailed object shapes from sparse views. Achieving optimal performance with these models can be challenging due to the extreme sparsity of inputs and distribution shifts induced by data corruptions. To this end, large, noise-free synthetic datasets can serve as shape priors to help models fill in gaps, but the resulting reconstructions must be approached with caution. Uncertainty estimation is crucial for assessing the quality of these reconstructions, particularly in identifying areas where the model is uncertain about the parts it has inferred from the prior. In this paper, we introduce Dropsembles, a novel method for uncertainty estimation in tuned

implicit functions. We demonstrate the efficacy of our approach through a series of experiments, starting with toy examples and progressing to a real-world scenario. Specifically, we train a Convolutional Occupancy Network on synthetic anatomical data and test it on low-resolution MRI segmentations of the lumbar spine. Our results show that Dropsembles achieve the accuracy and calibration levels of deep ensembles but with significantly less computational cost.

2899. On the Fourier analysis in the SO(3) space : the EquiLoPO Network

链接：https://iclr.cc/virtual/2025/poster/29964 abstract：

2900. Modeling Future Conversation Turns to Teach LLMs to Ask Clarifying Questions

链接：https://iclr.cc/virtual/2025/poster/29021 abstract： Large language models (LLMs) must often respond to highly ambiguous user requests. In such cases, the LLM's best response may be to ask a clarifying question to elicit more information. Existing LLMs often respond by presupposing a single interpretation of such ambiguous requests, frustrating users who intended a different interpretation. We speculate this is caused by current preference data labeling practice, where LLM responses are evaluated only on their prior contexts. To address this, we assign preference labels by simulating their expected outcomes in future turns. This allows LLMs to learn to ask clarifying questions when it can generate responses that are tailored to each user interpretation in future turns. On open-domain QA datasets with multiple annotations, we evaluate systems based on their ability to ask clarifying questions to recover each user's interpretation and expected answer. We compare systems trained using our proposed preference labeling methods against standard methods, which assign preferences based on only prior context. Our method achieves a 5% improvement in F1 measured against the answer set from different interpretations of each query, showing the value of modeling future conversation turns. We further demonstrate that our method can be used to train models to judiciously determine when to ask clarifying questions, directly answering the question when clarification is unnecessary. In our experiments, we find that our method achives a 3% improvement in accuracy of such judgments over existing methods.

2901. HaDeMiF: Hallucination Detection and Mitigation in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29391 abstract： The phenomenon of knowledge hallucinations has raised substantial concerns about the security and reliability of deployed large language models (LLMs). Current methods for detecting hallucinations primarily depend on manually designed individual metrics, such as prediction uncertainty and consistency, and fall short in effectively calibrating model predictions, thus constraining their detection accuracy and applicability in practical applications. In response, we propose an advanced framework, termed HaDeMiF, for detecting and mitigating hallucinations in LLMs. Specifically, hallucinations within the output and semantic spaces of LLMs are comprehensively captured through two compact networks—a novel, interpretable tree model known as the Deep Dynamic Decision Tree (D3T) and a Multilayer Perceptron (MLP)—which take as input a set of prediction characteristics and the hidden states of tokens, respectively. The predictions of LLMs are subsequently calibrated using the outputs from the D3T and MLP networks, aiming to mitigate hallucinations and enhance model calibration. HaDeMiF can be applied during both the inference and fine-tuning phases of LLMs, introducing less than 2% of the parameters relative to the LLMs through the training of two small-scale networks. Extensive experiments conclusively demonstrate the effectiveness of our framework in hallucination detection and model calibration across text generation tasks with responses of varying lengths.

2902. Meissonic: Revitalizing Masked Generative Transformers for Efficient High-Resolution Text-to-Image Synthesis

链接：https://iclr.cc/virtual/2025/poster/30289 abstract： We present Meissonic, which elevates non-autoregressive text-to-image Masked Image Modeling (MIM) to a level comparable with state-of-the-art diffusion models like SDXL. By incorporating a comprehensive suite of architectural innovations, advanced positional encoding strategies, and optimized sampling conditions, Meissonic substantially improves MIM's performance and efficiency. Additionally, we leverage high-quality training data, integrate micro-conditions informed by human preference scores, and employ feature compression layers to further enhance image fidelity and resolution. Our model not only matches but often exceeds the performance of existing methods in generating high-quality, high-resolution images. Extensive experiments validate Meissonic’s capabilities, demonstrating its potential as a new standard in text-to-image synthesis.

2903. Guided Score identity Distillation for Data-Free One-Step Text-to-Image Generation

链接：https://iclr.cc/virtual/2025/poster/30232 abstract： Diffusion-based text-to-image generation models trained on extensive text-image pairs have demonstrated the ability to produce photorealistic images aligned with textual descriptions. However, a significant limitation of these models is their slow sample generation process, which requires iterative refinement through the same network. To overcome this, we introduce a data-free guided distillation method that enables the efficient distillation of

pretrained Stable Diffusion models without access to the real training data, often restricted due to legal, privacy, or cost concerns. This method enhances Score identity Distillation (SiD) with Long and Short Classifier-Free Guidance (LSG), an innovative strategy that applies Classifier-Free Guidance (CFG) not only to the evaluation of the pretrained diffusion model but also to the training and evaluation of the fake score network. We optimize a model-based explicit score matching loss using a score-identity-based approximation alongside our proposed guidance strategies for practical computation. By exclusively training with synthetic images generated by its one-step generator, our data-free distillation method rapidly improves FID and CLIP scores, achieving state-of-the-art FID performance while maintaining a competitive CLIP score. Notably, the one-step distillation of Stable Diffusion 1.5 achieves an FID of 8.15 on the COCO-2014 validation set, a record low value under the data-free setting. Our code and checkpoints are available at https://github.com/mingyuanzhou/SiD-LSG.

2904. Adversarial Score identity Distillation: Rapidly Surpassing the Teacher in One Step

链接：https://iclr.cc/virtual/2025/poster/28519 abstract： Score identity Distillation (SiD) is a data-free method that has achieved state-of-the-art performance in image generation by leveraging only a pretrained diffusion model, without requiring any training data. However, the ultimate performance of SiD is constrained by the accuracy with which the pretrained model captures the true data scores at different stages of the diffusion process. In this paper, we introduce SiDA (SiD with Adversarial Loss), which not only enhances generation quality but also improves distillation efficiency by incorporating real images and adversarial loss. SiDA utilizes the encoder from the generator's score network as a discriminator, allowing it to distinguish between real images and those generated by SiD. The adversarial loss is batch-normalized within each GPU and then combined with the original SiD loss. This integration effectively incorporates the average "fakeness" per GPU batch into the pixel-based SiD loss, enabling SiDA to distill a single-step generator. SiDA converges significantly faster than its predecessor when distilled from scratch, and swiftly improves upon the original model's performance during fine-tuning from a pre-distilled SiD generator. This one-step adversarial distillation method establishes new benchmarks in generation performance when distilling EDM diffusion models, achieving FID scores of 1.499 on CIFAR-10 unconditional, 1.396 on CIFAR-10 conditional, and 1.110 on ImageNet 64x64. When distilling EDM2 models trained on ImageNet 512x512, our SiDA method surpasses even the largest teacher model, EDM2-XXL, which achieved an FID of 1.81 using classifier-free guidance (CFG) and 63 generation steps. Specifically, SiDA achieves FID scores of 2.156 for size XS, 1.669 for S, 1.488 for M, 1.413 for L, 1.379 for XL, and 1.366 for XXL, all without CFG and in a single generation step. These results highlight substantial improvements across all model sizes. Our code and checkpoints are available at https://github.com/mingyuanzhou/SiD/tree/sida.

2905. Adapting Multi-modal Large Language Model to Concept Drift From Pre-training Onwards

链接：https://iclr.cc/virtual/2025/poster/29128 abstract： Multi-modal Large Language Models (MLLMs) frequently face challenges from concept drift when dealing with real-world streaming data, wherein distributions change unpredictably. This mainly includes gradual drift due to long-tailed data and sudden drift from Out-Of-Distribution (OOD) data, both of which have increasingly drawn the attention of the research community. While these issues have been extensively studied in the individual domain of vision or language, their impacts on MLLMs in concept drift settings remain largely underexplored. In this paper, we reveal the susceptibility and vulnerability of Vision-Language (VL) models to significant biases arising from gradual drift and sudden drift, particularly in the pre-training. To effectively address these challenges, we propose a unified framework that extends concept drift theory to the multi-modal domain, enhancing the adaptability of the VL model to unpredictable distribution changes. Additionally, a T-distribution based drift adapter is proposed to effectively mitigate the bias induced by the gradual drift, which also facilitates the model in distinguishing sudden distribution changes through explicit distribution modeling. Extensive experiments demonstrate our method enhances the efficiency and accuracy of image-text alignment in the pre-training of VL models, particularly in the concept drift scenario. Moreover, various downstream tasks exhibit significant improvements in our model's ability to adapt to the long-tailed open world. Furthermore, we create a set of multi-modal datasets called OpenMMlo, specifically tailored for the long-tailed open-world setting, to validate our findings. To foster the development of the multi-modal community, we have made both OpenMMlo datasets and our code publicly available at: https://github.com/XiaoyuYoung/ConceptDriftMLLMs.

2906. Optimizing (L_0, L_1)-Smooth Functions by Gradient Methods

链接：https://iclr.cc/virtual/2025/poster/30281 abstract： We study gradient methods for optimizing (L_0, L_1)-smooth functions, aclass that generalizes Lipschitz-smooth functions and has gained attention forits relevance in machine learning.We provide new insights into the structure of this function class and developa principled framework for analyzing optimization methods in this setting.While our convergence rate estimates recover existing results for minimizingthe gradient norm in nonconvex problems, our approach significantly improvesthe best-known complexity bounds for convex objectives.Moreover, we show that the gradient method with Polyak stepsizes and thenormalized gradient method achieve nearly the same complexity guarantees asmethods that rely on explicit knowledge of (L_0, L_1).Finally, we demonstrate that a carefully designed accelerated gradientmethod can be applied to (L_0, L_1)-smooth functions, further improving allprevious results.

2907. cryoSPHERE: Single-Particle HEterogeneous REconstruction from cryo EM

链接：https://iclr.cc/virtual/2025/poster/28425 abstract： The three-dimensional structure of proteins plays a crucial role in determining their function. Protein structure prediction methods, like AlphaFold, offer rapid access to a protein’s structure. However, large protein complexes cannot be reliably predicted, and proteins are dynamic, making it important to resolve their full conformational distribution. Single-particle cryo-electron microscopy (cryo-EM) is a powerful tool for determining the structures of large protein complexes. Importantly, the numerous images of a given protein contain underutilized information about conformational heterogeneity. These images are very noisy projections of the protein, and traditional methods for cryo-EM reconstruction are limited to recovering only one or a few consensus conformations.In this paper, we introduce cryoSPHERE, which is a deep learning method that uses a nominal protein structure (e.g., from AlphaFold) as input, learns how to divide it into segments, and moves these segments as approximately rigid bodies to fit the different conformations present in the cryo-EM dataset. This approach provides enough constraints to enable meaningful reconstructions of single protein structural ensembles. We demonstrate this with two synthetic datasets featuring varying levels of noise, as well as two real dataset. We show that cryoSPHERE is very resilient to the high levels of noise typically encountered in experiments, where we see consistent improvements over the current state-of-the-art for heterogeneous reconstruction.

2908. Bridging the Gap between Variational Inference and Stochastic Gradient MCMC in Function Space

链接：https://iclr.cc/virtual/2025/poster/29112 abstract： Traditional parameter-space posterior inference for Bayesian neural networks faces several challenges, such as the difficulty in specifying meaningful prior, the potential pathologies in deep models and the intractability for multi-modal posterior. To address these issues, functional variational inference (fVI) and functional Markov Chain Monte Carlo (fMCMC) are two recently emerged Bayesian inference schemes that perform posterior inference directly in function space by incorporating more informative functional priors. Similar to their parameter-space counterparts, fVI and fMCMC have their own strengths and weaknesses. For instance, fVI is computationally efficient but imposes strong distributional assumptions, while fMCMC is asymptotically exact but suffers from slow mixing in high dimensions. To inherit the complementary benefits of both schemes, this work proposes a novel hybrid inference method for functional posterior inference. Specifically, it combines fVI and fMCMC successively by an elaborate linking mechanism to form an alternating approximation process. We also provide theoretical justification for the soundness of such a hybrid inference through the lens of Wasserstein gradient flows in the function space. We evaluate our method on several benchmark tasks and observe improvements in both predictive accuracy and uncertainty quantification compared to parameter/function-space VI and MCMC.

2909. Grammar Reinforcement Learning: path and cycle counting in graphs with a Context-Free Grammar and Transformer approach

链接：https://iclr.cc/virtual/2025/poster/27743 abstract： This paper presents Grammar Reinforcement Learning (GRL), a reinforcement learning algorithm that uses Monte Carlo Tree Search (MCTS) and a transformer architecture that models a Pushdown Automaton (PDA) within a context-free grammar (CFG) framework. Taking as use case the problem of efficiently counting paths and cycles in graphs, a key challenge in network analysis, computer science, biology, and social sciences, GRL discovers new matrix-based formulas for path/cycle counting that improve computational efficiency by factors of two to six w.r.t state-of-the-art approaches. Our contributions include: (i) a framework for generating transformers that operate within a CFG, (ii) the development of GRL for optimizing formulas within grammatical structures, and (iii) the discovery of novel formulas for graph substructure counting, leading to significant computational improvements.

2910. DICE: Data Influence Cascade in Decentralized Learning

链接：https://iclr.cc/virtual/2025/poster/31136 abstract： Decentralized learning offers a promising approach to crowdsource data consumptions and computational workloads across geographically distributed compute interconnected through peer-to-peer networks, accommodating the exponentially increasing demands. However, proper incentives are still in absence, considerably discouraging participation. Our vision is that a fair incentive mechanism relies on fair attribution of contributions to participating nodes, which faces non-trivial challenges arising from the localized connections making influence ``cascade" in a decentralized network. To overcome this, we design the first method to estimate Data Influence CascadE (DICE) in a decentralized environment. Theoretically, the framework derives tractable approximations of influence cascade over arbitrary neighbor hops, suggesting the influence cascade is determined by an interplay of data, communication topology, and the curvature of loss landscape.DICE also lays the foundations for applications including selecting suitable collaborators and identifying malicious behaviors.Project page is available at https://raiden-zhu.github.io/blog/2025/DICE.

2911. HyperFace: Generating Synthetic Face Recognition Datasets by Exploring Face Embedding Hypersphere

链接：https://iclr.cc/virtual/2025/poster/31005 abstract： Face recognition datasets are often collected by crawling Internet and without individuals' consents, raising ethical and privacy concerns. Generating synthetic datasets for training face recognition models has emerged as a promising alternative. However, the generation of synthetic datasets remains challenging as it entails adequate inter-class and intra-class variations. While advances in generative models have made it easier to increase intra-class variations in face datasets (such as pose, illumination, etc.), generating sufficient inter-class variation is still a difficult task. In this paper, we formulate the dataset generation as a packing problem on the embedding space (represented on a hypersphere) of a

face recognition model and propose a new synthetic dataset generation approach, called HyperFace. We formalize our packing problem as an optimization problem and solve it with a gradient descent-based approach. Then, we use a conditional face generator model to synthesize face images from the optimized embeddings. We use our generated datasets to train face recognition models and evaluate the trained models on several benchmarking real datasets. Our experimental results show that models trained with HyperFace achieve state-of-the-art performance in training face recognition using synthetic datasets. Project page: https://www.idiap.ch/paper/hyperface

2912. Rare-to-Frequent: Unlocking Compositional Generation Power of Diffusion Models on Rare Concepts with LLM Guidance

链接：https://iclr.cc/virtual/2025/poster/30557 abstract： State-of-the-art text-to-image (T2I) diffusion models often struggle to generate rare compositions of concepts, e.g., objects with unusual attributes. In this paper, we show that the compositional generation power of diffusion models on such rare concepts can be significantly enhanced by the Large Language Model (LLM) guidance. We start with empirical and theoretical analysis, demonstrating that exposing frequent concepts relevant to the target rare concepts during the diffusion sampling process yields more accurate concept composition. Based on this, we propose a training-free approach, R2F, that plans and executes the overall rare-to-frequent concept guidance throughout the diffusion inference by leveraging the abundant semantic knowledge in LLMs. Our framework is flexible across any pre-trained diffusion models and LLMs, and can be seamlessly integrated with the region-guided diffusion approaches. Extensive experiments on three datasets, including our newly proposed benchmark, RareBench, containing various prompts with rare compositions of concepts, R2F significantly surpasses existing models including SD3.0 and FLUX by up to 28.1%p in T2I alignment. Code is available at https://github.com/krafton-ai/Rare-to-Frequent.

2913. Release the Powers of Prompt Tuning: Cross-Modality Prompt Transfer

链接：https://iclr.cc/virtual/2025/poster/29594 abstract： Prompt Tuning adapts frozen models to new tasks by prepending a few learnable embeddings to the input.However, it struggles with tasks that suffer from data scarcity.To address this, we explore Cross-Modality Prompt Transfer, leveraging prompts pretrained on a data-rich modality to improve performance on data-scarce tasks in another modality.As a pioneering study, we first verify the feasibility of cross-modality prompt transfer by directly applying frozen source prompts (trained on the source modality) to the target modality task.To empirically study cross-modality prompt transferability, we train a linear layer to adapt source prompts to the target modality, thereby boosting performance and providing ground-truth transfer results.Regarding estimating prompt transferability, existing methods show ineffectiveness in cross-modality scenarios where the gap between source and target tasks is larger.We address this by decomposing the gap into the modality gap and the task gap, which we measure separately to autonomously select the best source prompt for a target task.Additionally, we propose Attention Transfer to further reduce the gaps by injecting target knowledge into the prompt and reorganizing a top-transferable source prompt using an attention block.We conduct extensive experiments involving prompt transfer from 13 source language tasks to 19 target vision tasks under three settings.Our findings demonstrate that:(i) cross-modality prompt transfer is feasible, supported by in-depth analysis;(ii) measuring both the modality and task gaps is crucial for accurate prompt transferability estimation, a factor overlooked by previous studies;(iii) cross-modality prompt transfer can significantly release the powers of prompt tuning on data-scarce tasks, as evidenced by comparisons with a newly released prompt-based benchmark.

2914. HMoRA: Making LLMs More Effective with Hierarchical Mixture of LoRA Experts

链接：https://iclr.cc/virtual/2025/poster/28518 abstract： Recent studies have combined Mixture of Experts (MoE) and Parameter-Efficient Fine-tuning (PEFT) to fine-tune large language models (LLMs), holding excellent performance in multi-task scenarios while remaining resource-efficient. However, existing MoE approaches still exhibit the following limitations: (1) Current methods fail to consider that different LLM layers capture features at varying levels of granularity, leading to suboptimal performance. (2) Task-level routing methods lack generalizability to unseen tasks. (3) The uncertainty introduced by load imbalance loss undermines the effective specialization of the experts. To address these challenges, we propose HMoRA, a Hierarchical fine-tuning method that combines MoE and LoRA, employing hybrid routing that integrates token-level and task-level routing in a hierarchical manner. This hierarchical hybrid routing allows the model to more efficiently capture both fine-grained token information and broader task contexts. To improve the certainty of expert selection, a novel routing auxiliary loss is introduced. This auxiliary function also enhances the task router's ability to differentiate tasks and its generalization to unseen tasks. Additionally, several optional lightweight designs have been proposed to significantly reduce both the number of trainable parameters and computational costs. Experimental results demonstrate that HMoRA outperforms full fine-tuning across multiple NLP benchmarks, while fine-tuning only 3.9\% of the parameters. The code is available on: https://github.com/LiaoMengqi/HMoRA.

2915. Navigating the Digital World as Humans Do: Universal Visual Grounding for GUI Agents

链接：https://iclr.cc/virtual/2025/poster/32062 abstract： Multimodal large language models (MLLMs) are transforming the

capabilities of graphical user interface (GUI) agents, facilitating their transition from controlled simulations to complex, real-world applications across various platforms. However, the effectiveness of these agents hinges on the robustness of their grounding capability. Current GUI agents predominantly utilize text-based representations such as HTML or accessibility trees, which, despite their utility, often introduce noise, incompleteness, and increased computational overhead. In this paper, we advocate a human-like embodiment for GUI agents that perceive the environment entirely visually and directly perform pixel-level operations on the GUI. The key is visual grounding models that can accurately map diverse referring expressions of GUI elements to their coordinates on the GUI across different platforms. We show that a simple recipe, which includes web-based synthetic data and slight adaptation of the LLaVA architecture, is surprisingly effective for training such visual grounding models. We collect the largest dataset for GUI visual grounding so far, containing 10M GUI elements and their referring expressions over 1.3M screenshots, and use it to train UGround, a strong universal visual grounding model for GUI agents. Empirical results on six benchmarks spanning three categories (grounding, offline agent, and online agent) show that 1) UGround substantially outperforms existing visual grounding models for GUI agents, by up to 20\% absolute, and 2) agents with UGround outperform state-of-the-art agents, despite the fact that existing agents use additional text-based input while ours only uses visual perception. These results provide strong support for the feasibility and promises of GUI agents that navigate the digital world as humans do.

2916. Going Beyond Static: Understanding Shifts with Time-Series Attribution

链接：https://iclr.cc/virtual/2025/poster/29306 abstract： Distribution shifts in time-series data are complex due to temporal dependencies, multivariable interactions, and trend changes. However, robust methods often rely on structural assumptions that lack thorough empirical validation, limiting their practical applicability. In order to support an empirically grounded inductive approach to research, we introduce our Time-Series Shift Attribution (TSSA) framework, which analyzes problem-specific patterns of distribution shifts. Our framework attributes performance degradation from various types of shifts to each temporal data property in a detailed manner, supported by theoretical analysis of unbiasedness and asymptotic properties. Empirical studies in real-world healthcare applications highlight how the TSSA framework enhances the understanding of time-series shifts, facilitating reliable model deployment and driving targeted improvements from both algorithmic and data-centric perspectives.

2917. Scrutinize What We Ignore: Reining In Task Representation Shift Of Context-Based Offline Meta Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/30493 abstract： Offline meta reinforcement learning (OMRL) has emerged as a promising approach for interaction avoidance and strong generalization performance by leveraging pre-collected data and meta-learning techniques. Previous context-based approaches predominantly rely on the intuition that alternating optimization between the context encoder and the policy can lead to performance improvements, as long as the context encoder follows the principle of maximizing the mutual information between the task variable M and its latent representation Z ($I(Z;M)$) while the policy adopts the standard offline reinforcement learning (RL) algorithms conditioning on the learned task representation.Despite promising results, the theoretical justification of performance improvements for such intuition remains underexplored.Inspired by the return discrepancy scheme in the model-based RL field, we find that the previous optimization framework can be linked with the general RL objective of maximizing the expected return, thereby explaining performance improvements. Furthermore, after scrutinizing this optimization framework, we observe that the condition for monotonic performance improvements does not consider the variation of the task representation. When these variations are considered, the previously established condition may no longer be sufficient to ensure monotonicity, thereby impairing the optimization process.We name this issue \underline{task representation shift} and theoretically prove that the monotonic performance improvements can be guaranteed with appropriate context encoder updates.We use different settings to rein in the task representation shift on three widely adopted training objectives concerning maximizing $I(Z;M)$ across different data qualities.Empirical results show that reining in the task representation shift can indeed improve performance.Our work opens up a new avenue for OMRL, leading to a better understanding between the task representation and performance improvements.

2918. Confidence Elicitation: A New Attack Vector for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29166 abstract： A fundamental issue in deep learning has been adversarial robustness. As these systems have scaled, such issues have persisted. Currently, large language models (LLMs) with billions of parameters suffer from adversarial attacks just like their earlier, smaller counterparts. However, the threat models have changed. Previously, having gray-box access, where input embeddings or output logits/probabilities were visible to the user, might have been reasonable. However, with the introduction of closed-source models, no information about the model is available apart from the generated output. This means that current black-box attacks can only utilize the final prediction to detect if an attack is successful. In this work, we investigate and demonstrate the potential of attack guidance, akin to using output probabilities, while having only black-box access in a classification setting. This is achieved through the ability to elicit confidence from the model. We empirically show that the elicited confidence is calibrated and not hallucinated for current LLMs. By minimizing the elicited confidence, we can therefore increase the likelihood of misclassification. Our new proposed paradigm demonstrates promising state-of-the-art results on three datasets across two models (LLaMA-3-8B-Instruct and Mistral-7B-Instruct-V0.3) when comparing our technique to existing hard-label black-box attack methods that introduce word-level substitutions. The code is publicly

available at GitHub: ConfidenceElicitationAttacks.

2919. PIED: Physics-Informed Experimental Design for Inverse Problems

链接：https://iclr.cc/virtual/2025/poster/27871 abstract： In many science and engineering settings, system dynamics are characterized by governing partial differential equations (PDEs), and a major challenge is to solve inverse problems (IPs) where unknown PDE parameters are inferred based on observational data gathered under limited budget. Due to the high costs of setting up and running experiments, experimental design (ED) is often done with the help of PDE simulations to optimize for the most informative design parameters (e.g., sensor placements) to solve such IPs, prior to actual data collection. This process of optimizing design parameters is especially critical when the budget and other practical constraints make it infeasible to adjust the design parameters between trials during the experiments.However, existing experimental design (ED) methods tend to require sequential and frequent design parameter adjustments between trials. Furthermore, they also have significant computational bottlenecks due to the need for complex numerical simulations for PDEs, and do not exploit the advantages provided by physics informed neural networks (PINNs) in solving IPs for PDE-governed systems, such as its meshless solutions, differentiability, and amortized training. This work presents Physics-Informed Experimental Design (PIED), the first ED framework that makes use of PINNs in a fully differentiable architecture to perform continuous optimization of design parameters for IPs for one-shot deployments. PIED overcomes existing methods' computational bottlenecks through parallelized computation and meta-learning of PINN parameter initialization, and proposes novel methods to effectively take into account PINN training dynamics in optimizing the ED parameters. Through experiments based on noisy simulated data and even real world experimental data, we empirically show that given limited observation budget, PIED significantly outperforms existing ED methods in solving IPs, including for challenging settings where the inverse parameters are unknown functions rather than just finite-dimensional.

2920. RelitLRM: Generative Relightable Radiance for Large Reconstruction Models

链接：https://iclr.cc/virtual/2025/poster/31081 abstract： We propose RelitLRM, a Large Reconstruction Model (LRM) for generating high-quality Gaussian splatting representations of 3D objects under novel illuminations from sparse (4-8) posed images captured under unknown static lighting. Unlike prior inverse rendering methods requiring dense captures and slow optimization, often causing artifacts like incorrect highlights or shadow baking, RelitLRM adopts a feed-forward transformer-based model with a novel combination of a geometry reconstructor and a relightable appearance generator based on diffusion. The model is trained end-to-end on synthetic multi-view renderings of objects under varying known illuminations. This architecture design enables to effectively decompose geometry and appearance, resolve the ambiguity between material and lighting, and capture the multi-modal distribution of shadows and specularity in the relit appearance. We show our sparse-view feed-forward RelitLRM offers competitive relighting results to state-of-the-art dense-view optimization-based baselines while being significantly faster. Our project page is available at: https://relit-lrm.github.io/.

2921. BIRD: A Trustworthy Bayesian Inference Framework for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28889 abstract： Predictive models often need to work with incomplete information in real-world tasks. Consequently, they must provide reliable probability or confidence estimation, especially in large-scale decision-making and planning tasks. Current large language models (LLMs) are insufficient for accurate estimations, but they can generate relevant factors that may affect the probabilities, produce coarse-grained probabilities when the information is more complete, and help determine which factors are relevant to specific downstream contexts. In this paper, we make use of these capabilities of LLMs to provide a significantly more accurate probabilistic estimation. We propose BIRD, a novel probabilistic inference framework that aligns a Bayesian network with LLM abductions and then estimates more accurate probabilities in a deduction step. We show BIRD provides reliable probability estimations that are 30% better than those provided directly by LLM baselines. These estimates further contribute to better and more trustworthy decision making.

2922. Decentralized Optimization with Coupled Constraints

链接：https://iclr.cc/virtual/2025/poster/30641 abstract： We consider the decentralized minimization of a separable objective $\sum_{i=1}^{n} f_i(x_i)$, where the variables are coupled through an affine constraint $\sum_{i=1}^n\left(\mathbf{A}_i x_i - b_i\right) = 0$.We assume that the functions f_i, matrices $\mathbf{A}_i$, and vectors b_i are stored locally by the nodes of a computational network, and that the functions f_i are smooth and strongly convex. This problem has significant applications in resource allocation and systems control and can also arise in distributed machine learning.We propose lower complexity bounds for decentralized optimization problems with coupled constraints and a first-order algorithm achieving the lower bounds. To the best of our knowledge, our method is also the first linearly convergent first-order decentralized algorithm for problems with general affine coupled constraints.

2923. Repulsive Latent Score Distillation for Solving Inverse Problems

链接：https://iclr.cc/virtual/2025/poster/29081 abstract： Score Distillation Sampling (SDS) has been pivotal for leveraging pre-

trained diffusion models in downstream tasks such as inverse problems, but it faces two major challenges: (i) mode collapse and (ii) latent space inversion, which become more pronounced in high-dimensional data. To address mode collapse, we introduce a novel variational framework for posterior sampling. Utilizing the Wasserstein gradient flow interpretation of SDS, we propose a multimodal variational approximation with a \emph{repulsion} mechanism that promotes diversity among particles by penalizing pairwise kernel-based similarity. This repulsion acts as a simple regularizer, encouraging a more diverse set of solutions. To mitigate latent space ambiguity, we extend this framework with an \emph{augmented} variational distribution that disentangles the latent and data. This repulsive augmented formulation balances computational efficiency, quality, and diversity. Extensive experiments on linear and nonlinear inverse tasks with high-resolution images (512×512) using pre-trained Stable Diffusion models demonstrate the effectiveness of our approach.

2924. A Visual Dive into Conditional Flow Matching

链接：https://iclr.cc/virtual/2025/poster/31356 abstract： Conditional flow matching was introduced by three simultaneous papers at ICLR 2023, through different approaches (conditional matching, rectifying flows and stochastic interpolants). In this blog post, we provide self-contained explanations and visualizations to understand standard flow techniques (Part 1) and conditional flow matching (Part 2). In addition we provide insights to grab new intuition on conditional flow matching (Part 3) .

2925. Youku Dense Caption: A Large-scale Chinese Video Dense Caption Dataset and Benchmarks

链接：https://iclr.cc/virtual/2025/poster/27881 abstract： With the explosive growth of video content, video captions have emerged as a crucial tool for video comprehension, significantly enhancing the ability to understand and retrieve information from videos. However, most publicly available dense video captioning datasets are in English, resulting in a scarcity of large-scale and high-quality Chinese dense video captioning datasets. To address this gap within the Chinese community and to promote the advancement of Chinese multi-modal models, we develop the first, large-scale, and high-quality Chinese dense video captioning dataset, named Youku Dense Caption. This dataset is sourced from Youku, a prominent Chinese video-sharing website. Youku Dense Caption includes 31,466 complete short videos annotated by 311,921 Chinese captions. To the best of our knowledge, it is currently the largest publicly available dataset for fine-grained Chinese video descriptions. Additionally, we establish several benchmarks for Chinese video-language tasks based on the Youku Dense Caption, including retrieval, grounding, and generation tasks. Extensive experiments and evaluations are conducted on existing state-of-the-art multi-modal models, demonstrating the dataset's utility and the potential for further research.

2926. Refine Knowledge of Large Language Models via Adaptive Contrastive Learning

链接：https://iclr.cc/virtual/2025/poster/30207 abstract： How to alleviate the hallucinations of Large Language Models (LLMs) has always been the fundamental goal pursued by the LLMs research community. Looking through numerous hallucination-related studies, a mainstream category of methods is to reduce hallucinations by optimizing the knowledge representation of LLMs to change their output. Considering that the core focus of these works is the knowledge acquired by models, and knowledge has long been a central theme in human societal progress, we believe that the process of models refining knowledge can greatly benefit from the way humans learn. In our work, by imitating the human learning process, we design an Adaptive Contrastive Learning strategy. Our method flexibly constructs different positive and negative samples for contrastive learning based on LLMs' actual mastery of knowledge. This strategy helps LLMs consolidate the correct knowledge they already possess, deepen their understanding of the correct knowledge they have encountered but not fully grasped, forget the incorrect knowledge they previously learned, and honestly acknowledge the knowledge they lack. Extensive experiments and detailed analyses on widely used datasets demonstrate the effectiveness and competitiveness of our method.

2927. Frame-Voyager: Learning to Query Frames for Video Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30004 abstract： Video Large Language Models (Video-LLMs) have made remarkable progress in video understanding tasks. However, they are constrained by the maximum length of input tokens, making it impractical to input entire videos. Existing frame selection approaches, such as uniform frame sampling and text-frame retrieval, fail to account for the information density variations in the videos or the complex instructions in the tasks, leading to sub-optimal performance. In this paper, we propose Frame-Voyager that learns to query informative frame combinations, based on the given textual queries in the task. To train Frame-Voyager, we introduce a new data collection and labeling pipeline, by ranking frame combinations using a pre-trained Video-LLM. Given a video of M frames, we traverse its T-frame combinations, feed them into a Video-LLM, and rank them based on Video-LLM's prediction losses. Using this ranking as supervision, we train Frame-Voyager to query the frame combinations with lower losses. In experiments, we evaluate Frame-Voyager on four Video Question Answering benchmarks by plugging it into two different Video-LLMs. The experimental results demonstrate that Frame-Voyager achieves impressive results in all settings, highlighting its potential as a plug-and-play solution for Video-LLMs.

2928. Zero-shot Imputation with Foundation Inference Models for Dynamical

Systems

链接：https://iclr.cc/virtual/2025/poster/29879 abstract： Dynamical systems governed by ordinary differential equations (ODEs) serve as models for a vast number of natural and social phenomena. In this work, we offer a fresh perspective on the classical problem of imputing missing time series data, whose underlying dynamics are assumed to be determined by ODEs. Specifically, we revisit ideas from amortized inference and neural operators, and propose a novel supervised learning framework for zero-shot time series imputation, through parametric functions satisfying some (hidden) ODEs. Our proposal consists of two components. First, a broad probability distribution over the space of ODE solutions, observation times and noise mechanisms, with which we generate a large, synthetic dataset of (hidden) ODE solutions, along with their noisy and sparse observations. Second, a neural recognition model that is trained offline, to map the generated time series onto the spaces of initial conditions and time derivatives of the (hidden) ODE solutions, which we then integrate to impute the missing data. We empirically demonstrate that one and the same (pretrained) recognition model can perform zero-shot imputation across 63 distinct time series with missing values, each sampled from widely different dynamical systems. Likewise, we demonstrate that it can perform zero-shot imputation of missing high-dimensional data in 10 vastly different settings, spanning human motion, air quality, traffic and electricity studies, as well as Navier-Stokes simulations — without requiring any fine-tuning. What is more, our proposal often outperforms state-of-the-art methods, which are trained on the target datasets.Our pretrained model, repository and tutorials are available online.

2929. Qinco2: Vector Compression and Search with Improved Implicit Neural Codebooks

链接：https://iclr.cc/virtual/2025/poster/31099 abstract： Vector quantization is a fundamental technique for compression and large-scale nearest neighbor search. For high-accuracy operating points, multi-codebook quantization associates data vectors with one element from each of multiple codebooks. An example is residual quantization (RQ), which iteratively quantizes the residual error of previous steps. Dependencies between the different parts of the code are, however, ignored in RQ, which leads to suboptimal rate-distortion performance. Qinco recently addressed this inefficiency by using a neural network to determine the quantization codebook in RQ based on the vector reconstruction from previous steps. In this paper we introduce Qinco2 which extends and improves Qinco with (i) improved vector encoding using codeword pre-selection and beam-search, (ii) a fast approximate decoder leveraging codeword pairs to establish accurate short-lists for search, and (iii) an optimized training procedure and network architecture. We conduct experiments on four datasets to evaluate Qinco2 for vector compression and billion-scale nearest neighbor search. We obtain outstanding results in both settings, improving the state-of-the-art reconstruction MSE by 44% for 16-byte vector compression on BigANN, and search accuracy by 24% with 8-byte encodings on Deep1M.

2930. Holistic Reasoning with Long-Context LMs: A Benchmark for Database Operations on Massive Textual Data

链接：https://iclr.cc/virtual/2025/poster/30950 abstract： The rapid increase in textual information means we need more efficient methods to sift through, organize, and understand it all. While retrieval-augmented generation (RAG) models excel in accessing information from large document collections, they struggle with complex tasks that require aggregation and reasoning over information spanning across multiple documents--what we call \textit{holistic reasoning}. Long-context language models (LCLMs) have great potential for managing large-scale documents, but their holistic reasoning capabilities remain unclear. In this work, we introduce HoloBench, a novel framework that brings database reasoning operations into text-based contexts, making it easier to systematically evaluate how LCLMs handle holistic reasoning across large documents. Our approach adjusts key factors such as context length, information density, distribution of information, and query complexity to evaluate LCLMs comprehensively. Our experiments show that the amount of information in the context has a bigger influence on LCLM performance than the actual context length. Furthermore, the complexity of queries affects performance more than the amount of information, particularly for different types of queries. Interestingly, queries that involve finding maximum or minimum values are easier for LCLMs and are less affected by context length, even though they pose challenges for RAG systems. However, tasks requiring the aggregation of multiple pieces of information show a noticeable drop in accuracy as context length increases. Additionally, we find that while grouping relevant information generally improves performance, the optimal positioning varies across models. Our findings surface both the advancements and the ongoing challenges in achieving a holistic understanding of long contexts. These can guide future developments in LCLMs and set the stage for creating more robust language models for real-world applications.

2931. Large Language Models Often Say One Thing and Do Another

链接：https://iclr.cc/virtual/2025/poster/29649 abstract： As large language models (LLMs) increasingly become central to various applications and interact with diverse user populations, ensuring their reliable and consistent performance is becoming more important. This paper explores a critical issue in assessing the reliability of LLMs: the consistency between their words and deeds. To quantitatively explore this consistency, we developed a novel evaluation benchmark called the Words and Deeds Consistency Test (WDCT). The benchmark establishes a strict correspondence between word-based and deed-based questions across different domains, including opinion vs. action, non-ethical value vs. action, ethical value vs. action, and theory vs. application. The evaluation results reveal a widespread inconsistency between words and deeds across different LLMs and domains. Subsequently, we conducted experiments with either word alignment or deed alignment to observe their impact on the

other aspect. The experimental results indicate that alignment only on words or deeds poorly and unpredictably influences the other aspect. This supports our hypothesis that the underlying knowledge guiding LLMs' word or deed choices is not contained within a unified space. Dataset and code are available at https://github.com/icip-cas/Word-Deed-Consistency-Test.

2932. Can a Large Language Model be a Gaslighter?

链接：https://iclr.cc/virtual/2025/poster/32085 abstract： Large language models (LLMs) have gained human trust due to their capabilities and helpfulness. However, this in turn may allow LLMs to affect users' mindsets by manipulating language. It is termed as gaslighting, a psychological effect. In this work, we aim to investigate the vulnerability of LLMs under prompt-based and fine-tuning-based gaslighting attacks. Therefore, we propose a two-stage framework DeepCoG designed to: 1) elicit gaslighting plans from LLMs with the proposed DeepGaslighting prompting template, and 2) acquire gaslighting conversations from LLMs through our Chain-of-Gaslighting method. The gaslighting conversation dataset along with a corresponding safe dataset is applied to fine-tuning-based attacks on open-source LLMs and anti-gaslighting safety alignment on these LLMs. Experiments demonstrate that both prompt-based and fine-tuning-based attacks transform three open-source LLMs into gaslighters. In contrast, we advanced three safety alignment strategies to strengthen~(by 12.05%) the safety guardrail of LLMs. Our safety alignment strategies have minimal impacts on the utility of LLMs. Empirical studies indicate that an LLM may be a potential gaslighter, even if it passed the harmfulness test on general dangerous queries.

2933. Deep Random Features for Scalable Interpolation of Spatiotemporal Data

链接：https://iclr.cc/virtual/2025/poster/29836 abstract： The rapid growth of earth observation systems calls for a scalable approach to interpolate remote-sensing observations. These methods in principle, should acquire more information about the observed field as data grows. Gaussian processes (GPs) are candidate model choices for interpolation. However, due to their poor scalability, they usually rely on inducing points for inference, which restricts their expressivity. Moreover, commonly imposed assumptions such as stationarity prevents them from capturing complex patterns in the data. While deep GPs can overcome this issue, training and making inference with them are difficult, again requiring crude approximations via inducing points. In this work, we instead approach the problem through Bayesian deep learning, where spatiotemporal fields are represented by deep neural networks, whose layers share the inductive bias of stationary GPs on the plane/sphere via random feature expansions. This allows one to (1) capture high frequency patterns in the data, and (2) use mini-batched gradient descent for large scale training. We experiment on various remote sensing data at local/global scales, showing that our approach produce competitive or superior results to existing methods, with well-calibrated uncertainties.

2934. Enhancing Vision-Language Model with Unmasked Token Alignment

链接：https://iclr.cc/virtual/2025/poster/31500 abstract： Contrastive pre-training on image-text pairs, exemplified by CLIP, becomes a standard technique for learning multi-modal visual-language representations. Although CLIP has demonstrated remarkable performance, training it from scratch on noisy web-scale datasets is computationally demanding. On the other hand, mask-then-predict pre-training approaches, like Masked Image Modeling (MIM), offer efficient self-supervised learning for single-modal representations. This paper introduces $\textbf{U}$nmasked $\textbf{T}$oken $\textbf{A}$lignment ($\textbf{UTA}$), a method that leverages existing CLIP models to further enhance its vision-language representations. UTA trains a Vision Transformer (ViT) by aligning unmasked visual tokens to the corresponding image tokens from a frozen CLIP vision encoder, which automatically aligns the ViT model with the CLIP text encoder. The pre-trained ViT can be directly applied for zero-shot evaluation even without training on image-text pairs. Compared to MIM approaches, UTA does not suffer from training-finetuning inconsistency and is much more training-efficient by avoiding using the extra $\mathrm{[MASK]}$ tokens. Extensive experimental results demonstrate that UTA can enhance CLIP models and outperform existing MIM methods on various uni- and multi-modal benchmarks.

2935. A3D: Does Diffusion Dream about 3D Alignment?

链接：https://iclr.cc/virtual/2025/poster/29688 abstract： We tackle the problem of text-driven 3D generation from a geometry alignment perspective. Given a set of text prompts, we aim to generate a collection of objects with semantically corresponding parts aligned across them. Recent methods based on Score Distillation have succeeded in distilling the knowledge from 2D diffusion models to high-quality representations of the 3D objects. These methods handle multiple text queries separately, and therefore the resulting objects have a high variability in object pose and structure. However, in some applications, such as 3D asset design, it may be desirable to obtain a set of objects aligned with each other. In order to achieve the alignment of the corresponding parts of the generated objects, we propose to embed these objects into a common latent space and optimize the continuous transitions between these objects. We enforce two kinds of properties of these transitions: smoothness of the transition and plausibility of the intermediate objects along the transition. We demonstrate that both of these properties are essential for good alignment. We provide several practical scenarios that benefit from alignment between the objects, including 3D editing and object hybridization, and experimentally demonstrate the effectiveness of our method.

2936. Input Space Mode Connectivity in Deep Neural Networks

链接：https://iclr.cc/virtual/2025/poster/31052 abstract： We extend the concept of loss landscape mode connectivity to the

input space of deep neural networks. Mode connectivity was originally studied within parameter space, where it describes the existence of low-loss paths between different solutions (loss minimizers) obtained through gradient descent. We present theoretical and empirical evidence of its presence in the input space of deep networks, thereby highlighting the broader nature of the phenomenon. We observe that different input images with similar predictions are generally connected, and for trained models, the path tends to be simple, with only a small deviation from being a linear path. Our methodology utilizes real, interpolated, and synthetic inputs created using the input optimization technique for feature visualization. We conjecture that input space mode connectivity in high-dimensional spaces is a geometric effect that takes place even in untrained models and can be explained through percolation theory. We exploit mode connectivity to obtain new insights about adversarial examples and demonstrate its potential for adversarial detection. Additionally, we discuss applications for the interpretability of deep networks.

2937. Making Transformer Decoders Better Differentiable Indexers

链接：https://iclr.cc/virtual/2025/poster/29100 abstract： Retrieval aims to find the top-k items most relevant to a query/user from a large dataset. Traditional retrieval models represent queries/users and items as embedding vectors and use Approximate Nearest Neighbor (ANN) search for retrieval. Recently, researchers have proposed a generative-based retrieval method that represents items as token sequences and uses a decoder model for autoregressive training. Compared to traditional methods, this approach uses more complex models and integrates index structure during training, leading to better performance. However, these methods remain two-stage processes, where index construction is separate from the retrieval model, limiting the model's overall capacity. Additionally, existing methods construct indices by clustering pre-trained item representations in Euclidean space. However, real-world scenarios are more complex, making this approach less accurate. To address these issues, we propose a \underline{U}nified framework for \underline{R}etrieval and \underline{I}ndexing, termed \textbf{URI}. URI ensures strong consistency between index construction and the retrieval model, typically a Transformer decoder. URI simultaneously builds the index and trains the decoder, constructing the index through the decoder itself. It no longer relies on one-sided item representations in Euclidean space but constructs the index within the interactive space between queries and items. Experimental comparisons on three real-world datasets show that URI significantly outperforms existing methods.

2938. What Makes Large Language Models Reason in (Multi-Turn) Code Generation?

链接：https://iclr.cc/virtual/2025/poster/29201 abstract： Prompting techniques such as chain-of-thought have established themselves as a popular vehicle for improving the outputs of large language models (LLMs). For code generation, however, their exact mechanics and efficacy are under-explored using unified metrics and benchmarks. We thus investigate the effects of a wide range of prompting strategies with a focus on automatic re-prompting over multiple turns and computational requirements. After systematically decomposing reasoning, instruction, and execution feedback prompts, we conduct an extensive grid search on the competitive programming benchmarks CodeContests and TACO for multiple LLM families and sizes (Llama 3.0 and 3.1, 8B, 70B, 405B, and GPT-4o). Our study reveals strategies that consistently improve performance across all models with small and large sampling budgets. We then show how finetuning with such an optimal configuration allows models to internalize the induced reasoning process and obtain improvements in performance and scalability for multi-turn code generation.

2939. P-SPIKESSM: HARNESSING PROBABILISTIC SPIKING STATE SPACE MODELS FOR LONG-RANGE DEPENDENCY TASKS

链接：https://iclr.cc/virtual/2025/poster/29587 abstract： Spiking neural networks (SNNs) are posited as a computationally efficient and biologically plausible alternative to conventional neural architectures, with their core computational framework primarily using the leaky integrate-and-fire (LIF) neuron model. However, the limited hidden state representation of LIF neurons, characterized by a scalar membrane potential, and sequential spike generation process, poses challenges for effectively developing scalable spiking models to address long-range dependencies in sequence learning tasks. In this study, we develop a scalable probabilistic spiking learning framework for long-range dependency tasks leveraging the fundamentals of state space models. Unlike LIF neurons that rely on the deterministic Heaviside function for a sequential process of spike generation, we introduce a SpikeSampler layer that samples spikes stochastically based on an SSM-based neuronal model while allowing parallel computations. To address non-differentiability of the spiking operation and enable effective training, we also propose a surrogate function tailored for the stochastic nature of the SpikeSampler layer. To enhance inter-neuron communication, we introduce the SpikeMixer block, which integrates spikes from neuron populations in each layer. This is followed by a ClampFuse layer, incorporating a residual connection to capture complex dependencies, enabling scalability of the model. Our models attain state-of-the-art performance among SNN models across diverse long-range dependency tasks, encompassing the Long Range Arena benchmark, permuted sequential MNIST, and the Speech Command dataset and demonstrate sparse spiking pattern highlighting its computational efficiency.

2940. The KoLMogorov Test: Compression by Code Generation

链接：https://iclr.cc/virtual/2025/poster/30532 abstract： Compression is at the heart of intelligence. A theoretically optimal way to compress any sequence of data is to find the shortest program that outputs that sequence and then halts. However, such Kolmogorov compression is uncomputable, and code generating LLMs struggle to approximate this theoretical ideal, as it requires reasoning, planning and search capabilities beyond those of current models. In this work, we introduce the

KoLMogorov-Test (KT), a compression-as-intelligence intelligence test for code generation LLMs. In KT a model is presented with a sequence of data at inference time, and asked to generate the shortest program that produces the sequence. We identify several benefits of KT for both evaluation and training: an essentially infinite number of problem instances of varying difficulty is readily available, strong baselines already exist, the evaluation metric (compression) cannot be gamed, and pretraining data contamination is highly unlikely. To evaluate current models, we use audio, text, and DNA data, as well as sequences produced by random synthetic programs. Current flagship models perform poorly - both GPT4-o and Llama-3.1-405B struggle on our natural and synthetic sequences. On our synthetic distribution, we are able to train code generation models with lower compression rates than previous approaches. Moreover, we show that gains on synthetic data generalize poorly to real data, suggesting that new innovations are necessary for additional gains on KT.

2941. Dynamic Mixture of Experts: An Auto-Tuning Approach for Efficient Transformer Models

链接：https://iclr.cc/virtual/2025/poster/29566 abstract： The Sparse Mixture of Experts (SMoE) has been widely employed to enhance the efficiency of training and inference for Transformer-based foundational models, yielding promising results. However, the performance of SMoE heavily depends on the choice of hyper-parameters, such as the number of experts and the number of experts to be activated (referred to as top-k), resulting in significant computational overhead due to the extensive model training by searching over various hyper-parameter configurations. As a remedy, we introduce the Dynamic Mixture of Experts (DynMoE) technique. DynMoE incorporates (1) a novel gating method that enables each token to automatically determine the number of experts to activate. (2) An adaptive process automatically adjusts the number of experts during training. Extensive numerical results across Vision, Language, and Vision-Language tasks demonstrate the effectiveness of our approach to achieve competitive performance compared to GMoE for vision and language tasks, and MoE-LLaVA for vision-language tasks, while maintaining efficiency by activating fewer parameters. Our code is available at \url{https://github.com/LINs-lab/DynMoE}.

2942. General Scene Adaptation for Vision-and-Language Navigation

链接：https://iclr.cc/virtual/2025/poster/31116 abstract： Vision-and-Language Navigation (VLN) tasks mainly evaluate agents based on one-time execution of individual instructions across multiple environments, aiming to develop agents capable of functioning in any environment in a zero-shot manner. However, real-world navigation robots often operate in persistent environments with relatively consistent physical layouts, visual observations, and language styles from instructors. Such a gap in the task setting presents an opportunity to improve VLN agents by incorporating continuous adaptation to specific environments. To better reflect these real-world conditions, we introduce GSA-VLN (General Scene Adaptation for VLN), a novel task requiring agents to execute navigation instructions within a specific scene and simultaneously adapt to it for improved performance over time. To evaluate the proposed task, one has to address two challenges in existing VLN datasets: the lack of out-of-distribution (OOD) data, and the limited number and style diversity of instructions for each scene. Therefore, we propose a new dataset, GSA-R2R, which significantly expands the diversity and quantity of environments and instructions for the Room-to-Room (R2R) dataset to evaluate agent adaptability in both ID and OOD contexts. Furthermore, we design a three-stage instruction orchestration pipeline that leverages large language models (LLMs) to refine speaker-generated instructions and apply role-playing techniques to rephrase instructions into different speaking styles. This is motivated by the observation that each individual user often has consistent signatures or preferences in their instructions, taking the use case of home robotic assistants as an example. We conducted extensive experiments on GSA-R2R to thoroughly evaluate our dataset and benchmark various methods, revealing key factors enabling agents to adapt to specific environments. Based on our findings, we propose a novel method, Graph-Retained DUET (GR-DUET), which incorporates memory-based navigation graphs with an environment-specific training strategy, achieving state-of-the-art results on all GSA-R2R splits.

2943. How to Evaluate Reward Models for RLHF

链接：https://iclr.cc/virtual/2025/poster/29039 abstract： We introduce a new benchmark for reward models that quantifies their ability to produce strong language models through RLHF (Reinforcement Learning from Human Feedback).The gold-standard approach is to run a full RLHF training pipeline and directly probe downstream LLM performance.However, this process is prohibitively expensive.To address this, we build a predictive model of downstream LLM performance by evaluating the reward model on proxy tasks. These proxy tasks consist of a large-scale human preference and a verifiable correctness preference dataset, in which we measure 12 metrics across 12 domains.To investigate which reward model metrics are most correlated to gold-standard RLHF outcomes, we launch an end-to-end RLHF experiment on a large-scale crowd-sourced human preference platform to view real reward model downstream performance as ground truth. Ultimately, we compile our data and findings into Preference Proxy Evaluations (PPE), the first reward model benchmark explicitly linked to post-RLHF real-world human preference performance, which we opensource for public use and further development at https://github.com/lmarena/PPE.

2944. Long Context Compression with Activation Beacon

链接：https://iclr.cc/virtual/2025/poster/31191 abstract： Long context compression is a critical research problem due to its significance in reducing the high computational and memory costs associated with LLMs. In this paper, we propose Activation Beacon, a plug-in module for transformer-based LLMs that targets effective, efficient, and flexible compression of long contexts. To achieve this, our method introduces the following technical designs. 1) We directly compress the activations (i.e. keys and values at every layer), rather than leveraging soft prompts to relay information (which constitute a major bottleneck to encapsulate

the complex information within long contexts).2) We tailor the compression workflow, where each fine-grained input unit is progressively compressed, enabling high-quality compression and efficient computation during both training and inference. 3) We train the model through compression-based auto-regression, making full use of plain texts and instructional data to optimize the model's compression performance.4) During training, we randomly sample a compression ratio at each step, teaching the model to support a wide range of compression configurations. Extensive evaluations are conducted on various long-context tasks whose lengths (e.g., 128K) may far exceed the maximum training length (20K), such as document understanding, few-shot learning, and Needle-in-a-Haystack. Whilst existing methods struggle to handle these challenging tasks, Activation Beacon maintains a comparable performance to the uncompressed baseline across various scenarios, achieving a 2x acceleration in inference time and an 8x reduction of memory costs for KV cache.

2945. RouteLLM: Learning to Route LLMs from Preference Data

链接：https://iclr.cc/virtual/2025/poster/30737 abstract： Large language models (LLMs) excel at a wide range of tasks, but choosing the right model often involves balancing performance and cost. Powerful models offer better results but are expensive, while smaller models are more cost-effective but less capable. To address this trade-off, we introduce a training framework for learning efficient router models that dynamically select between a stronger and weaker LLM during inference. Our framework leverages human preference data and employs data augmentation techniques to enhance performance. Evaluations on public benchmarks show that our approach can reduce costs by over 2 times without sacrificing response quality. Moreover, our routers exhibit strong generalization capabilities, maintaining performance even when routing between LLMs not included in training. This highlights the potential of our framework to deliver cost-effective, high-performance LLM solutions.

2946. SpikeLLM: Scaling up Spiking Neural Network to Large Language Models via Saliency-based Spiking

链接：https://iclr.cc/virtual/2025/poster/29210 abstract： Recent advancements in large language models (LLMs) with billions of parameters have improved performance in various applications, but their inference processes demand significant energy and computational resources. In contrast, the human brain, with approximately 86 billion neurons, is much more energy-efficient than LLMs with similar parameters. Inspired by this, we redesign 7$\sim$70 billion parameter LLMs using bio-plausible spiking mechanisms, emulating the efficient behavior of the human brain. We propose the first spiking large language model, SpikeLLM. Coupled with the proposed model, two essential approaches are proposed to improve spike training efficiency: Generalized Integrate-and-Fire (GIF) neurons to compress spike length from T to $\frac{T}{L} \log_2 L$ bits, and an Optimal Brain Spiking framework to divide outlier channels and allocate different T for GIF neurons, which further compresses spike length to approximate log_2T bits. The necessity of spike-driven LLM is proved by comparison with quantized LLMs with similar operations. In the OmniQuant pipeline, SpikeLLM reduces 11.01\% WikiText2 perplexity and improves 2.55\% accuracy of common scene reasoning on a LLAMA-7B W4A4 model. In the GPTQ pipeline, SpikeLLM achieves direct additive in linear layers, significantly exceeding PB-LLMs. Our code is publicly available at https://github.com/Xingrun-Xing2/SpikeLLM.

2947. Making Text Embedders Few-Shot Learners

链接：https://iclr.cc/virtual/2025/poster/27833 abstract： Large language models (LLMs) with decoder-only architectures have demonstrated exceptional text-generation capabilities across a variety of tasks. Some researchers have also adapted these models for text representation tasks. However, in text representation tasks, these models often face performance degradation on unseen tasks. In-context learning (ICL), which leverages examples provided in the input context, enables LLMs to handle unseen tasks effectively. Inspired by this, we aim to fully utilize the inherent properties of LLMs to enhance text representation performance across different tasks through the ICL approach.In this paper, we introduce a simple yet effective training strategy, which significantly improves text representation capabilities. Unlike previous models that prepend task instructions to the text, our method randomly samples a varying number of examples during training, endowing the embedding model with in-context learning abilities while maintaining its zero-shot capabilities. This approach does not require additional data construction or modifications to the model architecture. On the contrary, we find that some popular modifications to the model, such as bidirectional attention, can degrade performance, undermining the inherent characteristics of LLMs. We have publicly released our method at this \href{https://github.com/FlagOpen/FlagEmbedding}{repo}.

2948. K-HALU: Multiple Answer Korean Hallucination Benchmark for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/32080 abstract： Recent researchers and companies have been developing large language models (LLMs) specifically designed for particular purposes and have achieved significant advancements in various natural language processing tasks. However, LLMs are still prone to generating hallucinations—results that are unfaithful or inconsistent with the given input. As a result, the need for datasets to evaluate and demonstrate the hallucination detection capabilities of LLMs is increasingly recognized. Nonetheless, the Korean NLP community lacks publicly available benchmark datasets demonstrating the faithfulness of knowledge-based information. Furthermore, the few existing datasets that evaluate hallucination are limited in their access to the entire dataset, restricting detailed analysis beyond simple scoring, and are based on translated English knowledge. To address these challenges, we introduce K-HALU, a Korean benchmark designed to evaluate LLMs' hallucination detection in Korean. This benchmark contains seven domains, considering the faithfulness of statements based on knowledge documents compiled from Korean news, magazines, and books. For more strict evaluation,

40% of the dataset is structured as multiple-answer questions, requiring models to select all possible correct answers from the given options. Our empirical results show that open-source LLMs still struggle with hallucination detection in Korean knowledge, emphasizing the need for a more detailed analysis of their limitations.

2949. MMTEB: Massive Multilingual Text Embedding Benchmark

链接：https://iclr.cc/virtual/2025/poster/27651 abstract： Text embeddings are typically evaluated on a narrow set of tasks, limited in terms of languages, domains, and task types. To circumvent this limitation and to provide a more comprehensive evaluation, we introduce the Massive Multilingual Text Embedding Benchmark (MMTEB) -- a large-scale community-driven initiative expanding MTEB to over 500 quality-controlled evaluation tasks across 1,000+ languages. MMTEB includes a wide range of challenging novel tasks such as instruction following, long-document retrieval, and code retrieval, and represents the largest multilingual collection of evaluation tasks for embedding models to date. We use this collection to construct multiple highly multilingual benchmarks. We evaluate a representative set of models on these benchmarks.Our findings indicate that, while LLM-based models can achieve state-of-the-art performance on a subset of languages, the best-performing publicly available model across languages is the notably smaller, multilingual-e5-large-instruct.Massive benchmarks often impose high computational demands, limiting accessibility, particularly for low-resource communities. To address this, we downsample tasks based on inter-task correlation (i.e., selecting only a diverse set of tasks) while preserving relative rankings.We further optimize tasks such as retrieval by sampling hard negatives, creating smaller but effective splits. These optimizations allow us to introduce benchmarks at a significantly lower computational cost. For instance, we introduce a new zero-shot English benchmark that maintains a similar ordering at a fraction of the cost.

2950. AutoUAD: Hyper-parameter Optimization for Unsupervised Anomaly Detection

链接：https://iclr.cc/virtual/2025/poster/30384 abstract： Unsupervised anomaly detection (UAD) has important applications in diverse fields such as manufacturing industry and medical diagnosis. In the past decades, although numerous insightful and effective UAD methods have been proposed, it remains a huge challenge to tune the hyper-parameters of each method and select the most appropriate method among many candidates for a specific dataset, due to the absence of labeled anomalies in the training phase of UAD methods and the high diversity of real datasets. In this work, we aim to address this challenge, so as to make UAD more practical and reliable. We propose two internal evaluation metrics, relative-top-median and expected-anomaly-gap, and one semi-internal evaluation metric, normalized pseudo discrepancy (NPD), as surrogate functions of the expected model performance on unseen test data. For instance, NPD measures the discrepancy between the anomaly scores of a validation set drawn from the training data and a validation set drawn from an isotropic Gaussian. NPD is simple and hyper-parameter-free and is able to compare different UAD methods, and its effectiveness is theoretically analyzed. We integrate the three metrics with Bayesian optimization to effectively optimize the hyper-parameters of UAD models. Extensive experiments on 38 datasets show the effectiveness of our methods.

2951. FlashMask: Efficient and Rich Mask Extension of FlashAttention

链接：https://iclr.cc/virtual/2025/poster/27841 abstract： The computational and memory demands of vanilla attention scale quadratically with the sequence length N, posing significant challenges for processing long sequences in Transformer models. FlashAttention alleviates these challenges by eliminating the $\mathcal{O}(N^2)$ memory dependency and reducing attention latency through IO-aware memory optimizations. However, its native support for certain attention mask types is limited, and it does not inherently accommodate more complex masking requirements. Previous approaches resort to using dense masks with $\mathcal{O}(N^2)$ memory complexity, leading to inefficiencies. In this paper, we propose \ours{}, an extension of FlashAttention that introduces a column-wise sparse representation of attention masks. This approach efficiently represents a wide range of mask types and facilitates the development of optimized kernel implementations. By adopting this novel representation, \ours{} achieves linear memory complexity $\mathcal{O}(N)$, making it suitable for modeling long-context sequences. Moreover, this representation enables kernel optimizations that eliminate unnecessary computations by leveraging sparsity in the attention mask, without sacrificing computational accuracy, resulting in higher computational efficiency. We evaluate \ours{}'s performance in fine-tuning and alignment training of LLMs such as SFT, LoRA, DPO, and RM. \ours{} achieves significant throughput improvements, with end-to-end speedups ranging from 1.65x to 3.22x compared to existing FlashAttention dense method. Additionally, our kernel-level comparisons demonstrate that \ours{} surpasses the latest counterpart, FlexAttention, by 12.1\% to 60.7\% in terms of kernel TFLOPs/s, achieving 37.8\% to 62.3\% of the theoretical maximum FLOPs/s on the A100 GPU. The code is open-sourced on PaddlePaddle\footnote{\url{https://github.com/PaddlePaddle/Paddle}} and integrated into PaddleNLP\footnote{\url{https://github.com/PaddlePaddle/PaddleNLP}}, supporting models with over 100 billion parameters for contexts extending up to 128K tokens.

2952. EgoExo-Gen: Ego-centric Video Prediction by Watching Exo-centric Videos

链接：https://iclr.cc/virtual/2025/poster/30773 abstract： Generating videos in the first-person perspective has broad application prospects in the field of augmented reality and embodied intelligence.In this work, we explore the cross-view video prediction task, where given an exo-centric video, the first frame of the corresponding ego-centric video, and textual instructions,

the goal is to generate future frames of the ego-centric video. Inspired by the notion that hand-object interactions (HOI) in ego-centric videos represent the primary intentions and actions of the current actor, we present EgoExo-Gen that explicitly models the hand-object dynamics for cross-view video prediction. EgoExo-Gen consists of two stages. First, we design a cross-view HOI mask prediction model that anticipates the HOI masks in future ego-frames by modeling the spatio-temporal ego-exo correspondence. Next, we employ a video diffusion model to predict future ego-frames using the first ego-frame and textual instructions, while incorporating the HOI masks as structural guidance to enhance prediction quality.To facilitate training, we develop a fully automated pipeline to generate pseudo HOI masks for both ego- and exo-videos by exploiting vision foundation models. Extensive experiments demonstrate that our proposed EgoExo-Gen achieves better prediction performance compared to previous video prediction models on the public Ego-Exo4D and H2O benchmark datasets, with the HOI masks significantly improving the generation of hands and interactive objects in the ego-centric videos.

2953. CipherPrune: Efficient and Scalable Private Transformer Inference

链接：https://iclr.cc/virtual/2025/poster/28463 abstract： Private Transformer inference using cryptographic protocols offers promising solutions for privacy-preserving machine learning; however, it still faces significant runtime overhead (efficiency issues) and challenges in handling long-token inputs (scalability issues). We observe that the Transformer's operational complexity scales quadratically with the number of input tokens, making it essential to reduce the input token length. Notably, each token varies in importance, and many inputs contain redundant tokens. Additionally, prior private inference methods that rely on high-degree polynomial approximations for non-linear activations are computationally expensive. Therefore, reducing the polynomial degree for less important tokens can significantly accelerate private inference. Building on these observations, we propose \textit{CipherPrune}, an efficient and scalable private inference framework that includes a secure encrypted token pruning protocol, a polynomial reduction protocol, and corresponding Transformer network optimizations. At the protocol level, encrypted token pruning adaptively removes unimportant tokens from encrypted inputs in a progressive, layer-wise manner. Additionally, encrypted polynomial reduction assigns lower-degree polynomials to less important tokens after pruning, enhancing efficiency without decryption. At the network level, we introduce protocol-aware network optimization via a gradient-based search to maximize pruning thresholds and polynomial reduction conditions while maintaining the desired accuracy. Our experiments demonstrate that CipherPrune reduces the execution overhead of private Transformer inference by approximately $6.1\times$ for 128-token inputs and $10.6\times$ for 512-token inputs, compared to previous methods, with only a marginal drop in accuracy. The code is publicly available at https://github.com/UCF-Lou-Lab-PET/cipher-prune-inference.

2954. OpenHands: An Open Platform for AI Software Developers as Generalist Agents

链接：https://iclr.cc/virtual/2025/poster/29831 abstract： Software is one of the most powerful tools that we humans have at our disposal; it allows a skilled programmer to interact with the world in complex and profound ways. At the same time, thanks to improvements in large language models (LLMs), there has also been a rapid development in AI agents that interact with and effect change in their surrounding environments. In this paper, we introduce OpenHands, a platform for the development of powerful and flexible AI agents that interact with the world in similar ways to a human developer: by writing code, interacting with a command line, and browsing the web. We describe how the platform allows for the implementation of new agents, utilization of various LLMs, safe interaction with sandboxed environments for code execution, and incorporation of evaluation benchmarks. Based on our currently incorporated benchmarks, we perform an evaluation of agents over 13 challenging tasks, including software engineering (e.g., SWE-Bench) and web browsing (e.g., WebArena), amongst others. Released under the permissive MIT license, OpenHands is a community project spanning academia and industry with more than 2K contributions from over 186 contributors in less than six months of development, and will improve going forward.

2955. DON’T STOP ME NOW: EMBEDDING BASED SCHEDULING FOR LLMS

链接：https://iclr.cc/virtual/2025/poster/30829 abstract： Efficient scheduling is crucial for interactive Large Language Model (LLM) applications, where low request completion time directly impacts user engagement. Size-based scheduling algorithms like Shortest Remaining Process Time (SRPT) aim to reduce average request completion time by leveraging known or estimated request sizes and allowing preemption by incoming jobs with shorter service times. However, two main challenges arise when applying size-based scheduling to LLM systems. First, accurately predicting output lengths from prompts is challenging and often resource-intensive, making it impractical for many systems. As a result, the state-of-the-art LLM systems default to first-come, first-served scheduling, which can lead to head-of-line blocking and reduced system efficiency. Second, preemption introduces extra memory overhead to LLM systems as they must maintain intermediate states for unfinished (preempted) requests.In this paper, we propose TRAIL, a method to obtain output predictions from the target LLM itself. After generating each output token, we recycle the embedding of its internal structure as input for a lightweight classifier that predicts the remaining length for each running request. Using these predictions, we propose a prediction-based SRPT variant with limited preemption designed to account for memory overhead in LLM systems. This variant allows preemption early in request execution when memory consumption is low but restricts preemption as requests approach completion to optimize resource utilization. On the theoretical side, we derive a closed-form formula for this SRPT variant in an M/G/1 queue model, which demonstrates its potential value. In our system, we implement this preemption policy alongside our embedding-based prediction method. Our refined predictions from layer embeddings achieve 2.66x lower mean absolute error compared to BERT predictions from sequence prompts. TRAIL achieves 1.66x to 2.01x lower mean latency on the Alpaca dataset and 1.76x to 24.07x lower mean time to the first token compared to the state-of-the-art serving system.

2956. EMMA: Empowering Multi-modal Mamba with Structural and Hierarchical Alignment

链接：https://iclr.cc/virtual/2025/poster/30381 abstract： Mamba-based architectures have shown to be a promising new direction for deep learning models owing to their competitive performance and sub-quadratic deployment speed. However, current Mamba multi-modal large language models (MLLM) are insufficient in extracting visual features, leading to imbalanced cross-modal alignment between visual and textural latents, negatively impacting performance on multi-modal tasks. In this work, we propose Empowering Multi-modal Mamba with Structural and Hierarchical Alignment (EMMA), which enables the MLLM to extract fine-grained visual information. Specifically, we propose a pixel-wise alignment module to autoregressively optimize the learning and processing of spatial image-level features along with textual tokens, enabling structural alignment at the image level. In addition, to prevent the degradation of visual information during the cross-model alignment process, we propose a multi-scale feature fusion (MFF) module to combine multi-scale visual features from intermediate layers, enabling hierarchical alignment at the feature level. Extensive experiments are conducted across a variety of multi-modal benchmarks. Our model shows lower latency than other Mamba-based MLLMs and is nearly four times faster than transformer-based MLLMs of similar scale during inference. Due to better cross-modal alignment, our model exhibits lower degrees of hallucination and enhanced sensitivity to visual details, which manifests in superior performance across diverse multi-modal benchmarks. Code provided at https://github.com/xingyifei2016/EMMA.

2957. FreCaS: Efficient Higher-Resolution Image Generation via Frequency-aware Cascaded Sampling

链接：https://iclr.cc/virtual/2025/poster/29507 abstract： While image generation with diffusion models has achieved a great success, generating images of higher resolution than the training size remains a challenging task due to the high computational cost. Current methods typically perform the entire sampling process at full resolution and process all frequency components simultaneously, contradicting with the inherent coarse-to-fine nature of latent diffusion models and wasting computations on processing premature high-frequency details at early diffusion stages. To address this issue, we introduce an efficient $\textbf{Fre}$quency-aware $\textbf{Ca}$scaded $\textbf{S}$ampling framework, $\textbf{FreCaS}$ in short, for higher-resolution image generation. FreCaS decomposes the sampling process into cascaded stages with gradually increased resolutions, progressively expanding frequency bands and refining the corresponding details. We propose an innovative frequency-aware classifier-free guidance (FA-CFG) strategy to assign different guidance strengths for different frequency components, directing the diffusion model to add new details in the expanded frequency domain of each stage. Additionally, we fuse the cross-attention maps of previous and current stages to avoid synthesizing unfaithful layouts. Experiments demonstrate that FreCaS significantly outperforms state-of-the-art methods in image quality and generation speed. In particular, FreCaS is about 2.86$\times$ and 6.07$\times$ faster than ScaleCrafter and DemoFusion in generating a 2048$\times$2048 image using a pretrained SDXL model and achieves an FID_b improvement of 11.6 and 3.7, respectively. FreCaS can be easily extended to more complex models such as SD3. The source code of FreCaS can be found at https://github.com/xtudbxk/FreCaS.

2958. How Much is Unseen Depends Chiefly on Information About the Seen

链接：https://iclr.cc/virtual/2025/poster/27955 abstract： The *missing mass* refers to the proportion of data points in an *unknown* population of classifier inputs that belong to classes *not* present in the classifier's training data, which is assumed to be a random sample from that unknown population.We find that *in expectation* the missing mass is entirely determined by the number f_k of classes that *do* appear in the training data the same number of times *and an exponentially decaying error*.While this is the first precise characterization of the expected missing mass in terms of the sample, the induced estimator suffers from an impractically high variance. However, our theory suggests a large search space of nearly unbiased estimators that can be searched effectively and efficiently. Hence, we cast distribution-free estimation as an optimization problem to find a distribution-specific estimator with a minimized mean-squared error (MSE), given only the sample.In our experiments, our search algorithm discovers estimators that have a substantially smaller MSE than the state-of-the-art Good-Turing estimator. This holds for over 93\% of runs when there are at least as many samples as classes. Our estimators' MSE is roughly 80\% of the Good-Turing estimator's.

2959. ZAPBench: A Benchmark for Whole-Brain Activity Prediction in Zebrafish

链接：https://iclr.cc/virtual/2025/poster/28372 abstract： Data-driven benchmarks have led to significant progress in key scientific modeling domains including weather and structural biology. Here, we introduce the Zebrafish Activity Prediction Benchmark (ZAPBench) to measure progress on the problem of predicting cellular-resolution neural activity throughout an entire vertebrate brain. The benchmark is based on a novel dataset containing 4d light-sheet microscopy recordings of over 70,000 neurons in a larval zebrafish brain, along with motion stabilized and voxel-level cell segmentations of these data that facilitate development of a variety of forecasting methods. Initial results from a selection of time series and volumetric video modeling approaches achieve better performance than naive baseline methods, but also show room for further improvement. The specific brain used in the activity recording is also undergoing synaptic-level anatomical mapping, which will enable future integration of detailed structural information into forecasting methods.

2960. Spatial-Mamba: Effective Visual State Space Models via Structure-Aware State Fusion

链接：https://iclr.cc/virtual/2025/poster/28708 abstract： Selective state space models (SSMs), such as Mamba, highly excel at capturing long-range dependencies in 1D sequential data, while their applications to 2D vision tasks still face challenges. Current visual SSMs often convert images into 1D sequences and employ various scanning patterns to incorporate local spatial dependencies. However, these methods are limited in effectively capturing the complex image spatial structures and the increased computational cost caused by the lengthened scanning paths. To address these limitations, we propose Spatial-Mamba, a novel approach that establishes neighborhood connectivity directly in the state space. Instead of relying solely on sequential state transitions, we introduce a structure-aware state fusion equation, which leverages dilated convolutions to capture image spatial structural dependencies, significantly enhancing the flow of visual contextual information. Spatial-Mamba proceeds in three stages: initial state computation in a unidirectional scan, spatial context acquisition through structure-aware state fusion, and final state computation using the observation equation. Our theoretical analysis shows that Spatial-Mamba unifies the original Mamba and linear attention under the same matrix multiplication framework, providing a deeper understanding of our method. Experimental results demonstrate that Spatial-Mamba, even with a single scan, attains or surpasses the state-of-the-art SSM-based models in image classification, detection and segmentation. Source codes and trained models can be found at \url{ https://github.com/EdwardChasel/Spatial-Mamba }.

2961. Loss Landscape of Shallow ReLU-like Neural Networks: Stationary Points, Saddle Escape, and Network Embedding

链接：https://iclr.cc/virtual/2025/poster/28339 abstract： In this paper, we study the loss landscape of one-hidden-layer neural networks with ReLU-like activation functions trained with the empirical squared loss using gradient descent (GD). We identify the stationary points of such networks, which significantly slow down loss decrease during training. To capture such points while accounting for the non-differentiability of the loss, the stationary points that we study are directional stationary points, rather than other notions like Clarke stationary points. We show that, if a stationary point does not contain "escape neurons", which are defined with first-order conditions, it must be a local minimum. Moreover, for the scalar-output case, the presence of an escape neuron guarantees that the stationary point is not a local minimum. Our results refine the description of the saddle-to-saddle training process starting from infinitesimally small (vanishing) initialization for shallow ReLU-like networks: By precluding the saddle escape types that previous works did not rule out, we advance one step closer to a complete picture of the entire dynamics. Moreover, we are also able to fully discuss how network embedding, which is to instantiate a narrower network with a wider network, reshapes the stationary points.

2962. Quantum-PEFT: Ultra parameter-efficient fine-tuning

链接：https://iclr.cc/virtual/2025/poster/28975 abstract： This paper introduces Quantum-PEFT that leverages quantum computations for parameter-efficient fine-tuning (PEFT). Unlike other additive PEFT methods, such as low-rank adaptation (LoRA), Quantum-PEFT exploits an underlying full-rank yet surprisingly parameter efficient quantum unitary parameterization. With the use of Pauli parameterization, the number of trainable parameters grows only logarithmically with the ambient dimension, as opposed to linearly as in LoRA-based PEFT methods. Quantum-PEFT achieves vanishingly smaller number of trainable parameters than the lowest-rank LoRA as dimensions grow, enhancing parameter efficiency while maintaining a competitive performance. We apply Quantum-PEFT to several transfer learning benchmarks in language and vision, demonstrating significant advantages in parameter efficiency.

2963. Data Selection via Optimal Control for Language Models

链接：https://iclr.cc/virtual/2025/poster/28972 abstract： This work investigates the selection of high-quality pre-training data from massive corpora to enhance LMs' capabilities for downstream usage. We formulate data selection as a generalized Optimal Control problem, which can be solved theoretically by Pontryagin's Maximum Principle (PMP), yielding a set of necessary conditions that characterize the relationship between optimal data selection and LM training dynamics.Based on these theoretical results, we introduce PMP-based Data Selection (PDS), a framework that approximates optimal data selection by solving the PMP conditions. In our experiments, we adopt PDS to select data from CommmonCrawl and show that the PDS-selected corpus accelerates the learning of LMs and constantly boosts their performance on a wide range of downstream tasks across various model sizes.Moreover, the benefits of PDS extend to ~400B models trained on ~10T tokens, as evidenced by the extrapolation of the test loss curves according to the Scaling Laws.PDS also improves data utilization when the pre-training data is limited, by reducing the data demand by 1.8 times, which helps mitigate the quick exhaustion of available web-crawled corpora. Our code, model, and data can be found at https://github.com/microsoft/LMOps/tree/main/data_selection.

2964. TeaserGen: Generating Teasers for Long Documentaries

链接：https://iclr.cc/virtual/2025/poster/32100 abstract： Teasers are an effective tool for promoting content in entertainment, commercial and educational fields. However, creating an effective teaser for long videos is challenging for it requires long-range multimodal modeling capability for the input videos, while necessitating maintaining audiovisual alignments, managing scene transitions and preserving factual accuracy for the output teasers. Due to the lack of a publicly-available dataset, progress along

this research direction has been hindered. In this work, we present DocumentaryNet, a collection of 1,269 documentaries paired with their teasers, featuring multimodal data streams of video, speech, music, sound effects and narrations. With DocumentaryNet, we propose a new two-stage system for generating teasers from long documentaries. The proposed TeaserGen system first generates the teaser narration from the transcribed narration from the documentary using a pretrained large language model, and then selects the most relevant visual content to accompany the generated narration through language-vision models. For narration-video matching, we explore two approaches: a pretraining-based model using pretrained contrastive language-vision models and a deep sequential model that learns the mapping between the narrations and visuals. Our experimental results show that the pretraining-based approach is more effective at identifying relevant visual content than directly trained deep autoregressive models.

2965. Aligning Visual Contrastive learning models via Preference Optimization

链接：https://iclr.cc/virtual/2025/poster/27829 abstract： Contrastive learning models have demonstrated impressive abilities to capture semantic similarities by aligning representations in the embedding space. However, their performance can be limited by the quality of the training data and its inherent biases. While Preference Optimization (PO) methods such as Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) have been applied to align generative models with human preferences, their use in contrastive learning has yet to be explored. This paper introduces a novel method for training contrastive learning models using different PO methods to break down complex concepts. Our method systematically aligns model behavior with desired preferences, enhancing performance on the targeted task. In particular, we focus on enhancing model robustness against typographic attacks and inductive biases, commonly seen in contrastive vision-language models like CLIP. Our experiments demonstrate that models trained using PO outperform standard contrastive learning techniques while retaining their ability to handle adversarial challenges and maintain accuracy on other downstream tasks. This makes our method well-suited for tasks requiring fairness, robustness, and alignment with specific preferences. We evaluate our method for tackling typographic attacks on images and explore its ability to disentangle gender concepts and mitigate gender bias, showcasing the versatility of our approach.

2966. VVC-Gym: A Fixed-Wing UAV Reinforcement Learning Environment for Multi-Goal Long-Horizon Problems

链接：https://iclr.cc/virtual/2025/poster/30922 abstract： Multi-goal long-horizon problems are prevalent in real-world applications. The additional goal space introduced by multi-goal problems intensifies the spatial complexity of exploration; meanwhile, the long interaction sequences in long-horizon problems exacerbate the temporal complexity of exploration. Addressing the great exploration challenge posed by multi-goal long-horizon problems depends not only on the design of algorithms but also on the design of environments and the availability of demonstrations to assist in training. To facilitate the above research, we propose a multi-goal long-horizon Reinforcement Learning (RL) environment based on realistic fixed-wing UAV's velocity vector control, named VVC-Gym, and generate multiple demonstration sets of various quality. Through experimentation, we analyze the impact of different environment designs on training, assess the quantity and quality of demonstrations and their influence on training, and assess the effectiveness of various RL algorithms, providing baselines on VVC-Gym and its corresponding demonstrations. The results suggest that VVC-Gym is suitable for studying: (1) the influence of environment designs on addressing multi-goal long-horizon problems with RL. (2) the assistance that demonstrations can provide in overcoming the exploration challenges of multi-goal long-horizon problems. (3) the RL algorithm designs with the least possible impact from environment designs on the efficiency and effectiveness of training.

2967. Exploring a Principled Framework for Deep Subspace Clustering

链接：https://iclr.cc/virtual/2025/poster/30796 abstract： Subspace clustering is a classical unsupervised learning task, built on a basic assumption that high-dimensional data can be approximated by a union of subspaces (UoS). Nevertheless, the real-world data are often deviating from the UoS assumption. To address this challenge, state-of-the-art deep subspace clustering algorithms attempt to jointly learn UoS representations and self-expressive coefficients. However, the general framework of the existing algorithms suffers from feature collapse and lacks a theoretical guarantee to learn desired UoS representation. In this paper, we present a Principled fRamewOrk for Deep Subspace Clustering (PRO-DSC), which is designed to learn structured representations and self-expressive coefficients in a unified manner. Specifically, in PRO-DSC, we incorporate an effective regularization on the learned representations into the self-expressive model, prove that the regularized self-expressive model is able to prevent feature space collapse, and demonstrate that the learned optimal representations under certain condition lie on a union of orthogonal subspaces. Moreover, we provide a scalable and efficient approach to implement our PRO-DSC and conduct extensive experiments to verify our theoretical findings and demonstrate the superior performance of our proposed deep subspace clustering approach.

2968. HERO: Human-Feedback Efficient Reinforcement Learning for Online Diffusion Model Finetuning

链接：https://iclr.cc/virtual/2025/poster/27733 abstract： Controllable generation through Stable Diffusion (SD) fine-tuning aims to improve fidelity, safety, and alignment with human guidance. Existing reinforcement learning from human feedback methods

usually rely on predefined heuristic reward functions or pretrained reward models built on large-scale datasets, limiting their applicability to scenarios where collecting such data is costly or difficult. To effectively and efficiently utilize human feedback, we develop a framework, HERO, which leverages online human feedback collected on the fly during model learning. Specifically, HERO features two key mechanisms: (1) Feedback-Aligned Representation Learning, an online training method that captures human feedback and provides informative learning signals for fine-tuning, and (2) Feedback-Guided Image Generation, which involves generating images from SD's refined initialization samples, enabling faster convergence towards the evaluator's intent. We demonstrate that HERO is 4x more efficient in online feedback for body part anomaly correction compared to the best existing method. Additionally, experiments show that HERO can effectively handle tasks like reasoning, counting, personalization, and reducing NSFW content with only 0.5K online feedback. The code and project page are available at https://hero-dm.github.io/.

2969. Scaling Laws for Downstream Task Performance in Machine Translation

链接：https://iclr.cc/virtual/2025/poster/27915 abstract： Scaling laws provide important insights that can guide the design of large language models (LLMs). Existing work has primarily focused on studying scaling laws for pretraining (upstream) loss. However, in transfer learning settings, in which LLMs are pretrained on an unsupervised dataset and then finetuned on a downstream task, we often also care about the downstream performance. In this work, we study the scaling behavior in a transfer learning setting, where LLMs are finetuned for machine translation tasks. Specifically, we investigate how the choice of the pretraining data and its size affect downstream performance (translation quality) as judged by: downstream cross-entropy and translation quality metrics such as BLEU and COMET scores. Our experiments indicate that the size of the finetuning dataset and the distribution alignment between the pretraining and downstream data significantly influence the scaling behavior. With sufficient alignment, both downstream cross-entropy and translation quality scores improve monotonically with more pretraining data. In such cases, we show that it is possible to predict the downstream translation quality metrics with good accuracy using a log-law. However, there are cases where moderate misalignment causes the downstream translation scores to fluctuate or get worse with more pretraining, whereas downstream cross-entropy monotonically improves. By analyzing these, we provide new practical insights for choosing appropriate pretraining data.

2970. Ranking-aware adapter for text-driven image ordering with CLIP

链接：https://iclr.cc/virtual/2025/poster/30043 abstract： Recent advances in vision-language models (VLMs) have made significant progress in downstream tasks that require quantitative concepts such as facial age estimation and image quality assessment, enabling VLMs to explore applications like image ranking and retrieval. However, existing studies typically focus on the reasoning based on a single image and heavily depend on text prompting, limiting their ability to learn comprehensive understanding from multiple images. To address this, we propose an effective yet efficient approach that reframes the CLIP model into a learning-to-rank task and introduces a lightweight adapter to augment CLIP for text-guided image ranking. Specifically, our approach incorporates learnable prompts to adapt to new instructions for ranking purposes and an auxiliary branch with ranking-aware attention, leveraging text-conditioned visual differences for additional supervision in image ranking. Our ranking-aware adapter consistently outperforms fine-tuned CLIPs on various tasks and achieves competitive results compared to state-of-the-art models designed for specific tasks like facial age estimation and image quality assessment. Overall, our approach primarily focuses on ranking images with a single instruction, which provides a natural and generalized way of learning from visual differences across images, bypassing the need for extensive text prompts tailored to individual tasks.

2971. Towards Optimal Multi-draft Speculative Decoding

链接：https://iclr.cc/virtual/2025/poster/30699 abstract： Large Language Models (LLMs) have become an indispensable part of natural language processing tasks. However, autoregressive sampling has become an efficiency bottleneck. Multi-Draft Speculative Decoding (MDSD) is a recent approach where, when generating each token, a small draft model generates multiple drafts, and the target LLM verifies them in parallel, ensuring that the final output conforms to the target model distribution. The two main design choices in MDSD are the draft sampling method and the verification algorithm. For a fixed draft sampling method, the optimal acceptance rate is a solution to an optimal transport problem, but the complexity of this problem makes it difficult to solve for the optimal acceptance rate and measure the gap between existing verification algorithms and the theoretical upper bound. This paper discusses the dual of the optimal transport problem, providing a way to efficiently compute the optimal acceptance rate. For the first time, we measure the theoretical upper bound of MDSD efficiency for vocabulary sizes in the thousands and quantify the gap between existing verification algorithms and this bound. We also compare different draft sampling methods based on their optimal acceptance rates. Our results show that the draft sampling method strongly influences the optimal acceptance rate, with sampling without replacement outperforming sampling with replacement. Additionally, existing verification algorithms do not reach the theoretical upper bound for both without replacement and with replacement sampling. Our findings suggest that carefully designed draft sampling methods can potentially improve the optimal acceptance rate and enable the development of verification algorithms that closely match the theoretical upper bound.

2972. On the Feature Learning in Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/30093 abstract： The predominant success of diffusion models in generative modeling has spurred significant interest in understanding their theoretical foundations. In this work, we propose a feature learning framework aimed at analyzing and comparing the training dynamics of diffusion models with those of traditional classification

models. Our theoretical analysis demonstrates that diffusion models, due to the denoising objective, are encouraged to learn more balanced and comprehensive representations of the data. In contrast, neural networks with a similar architecture trained for classification tend to prioritize learning specific patterns in the data, often focusing on easy-to-learn components. To support these theoretical insights, we conduct several experiments on both synthetic and real-world datasets, which empirically validate our findings and highlight the distinct feature learning dynamics in diffusion models compared to classification.

2973. Q-SFT: Q-Learning for Language Models via Supervised Fine-Tuning

链接：https://iclr.cc/virtual/2025/poster/27935 abstract： Value-based reinforcement learning (RL) can in principle learn effective policies for a wide range of multi-turn problems, from games to dialogue to robotic control, including via offline RL from static previously collected datasets. However, despite the widespread use of policy gradient methods to train large language models for single turn tasks (e.g., question answering), value-based methods for multi-turn RL in an off-policy or offline setting have proven particularly challenging to scale to the setting of large language models. This setting requires effectively leveraging pretraining, scaling to large architectures with billions of parameters, and training on large datasets, all of which represent major challenges for current value-based RL methods. In this work, we propose a novel offline RL algorithm that addresses these drawbacks, casting Q-learning as a modified supervised fine-tuning (SFT) problem where the probabilities of tokens directly translate to Q-values. In this way we obtain an algorithm that smoothly transitions from maximizing the likelihood of the data during pretraining to learning a near-optimal Q-function during finetuning. Our algorithm has strong theoretical foundations, enjoying performance bounds similar to state-of-the-art Q-learning methods, while in practice utilizing an objective that closely resembles SFT. Because of this, our approach can enjoy the full benefits of the pretraining of language models, without the need to reinitialize any weights before RL finetuning, and without the need to initialize new heads for predicting values or advantages. Empirically, we evaluate our method on both pretrained LLMs and VLMs, on a variety of tasks including both natural language dialogue and robotic manipulation and navigation from images.

2974. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer

链接：https://iclr.cc/virtual/2025/poster/29998 abstract： We present CogVideoX, a large-scale text-to-video generation model based on diffusion transformer, which can generate 10-second continuous videos that align seamlessly with text prompts, with a frame rate of 16 fps and resolution of 768 x 1360 pixels. Previous video generation models often struggled with limited motion and short durations.It is especially difficult to generate videos with coherent narratives based on text. We propose several designs to address these issues. First, we introduce a 3D Variational Autoencoder (VAE) to compress videos across spatial and temporal dimensions, enhancing both the compression rate and video fidelity. Second, to improve text-video alignment, we propose an expert transformer with expert adaptive LayerNorm to facilitate the deep fusion between the two modalities.Third, by employing progressive training and multi-resolution frame packing, CogVideoX excels at generating coherent, long-duration videos with diverse shapes and dynamic movements. In addition, we develop an effective pipeline that includes various pre-processing strategies for text and video data.Our innovative video captioning model significantly improves generation quality and semantic alignment. Results show that CogVideoX achieves state-of-the-art performance in both automated benchmarks and human evaluation.We publish the code and model checkpoints of CogVideoX along with our VAE model and video captioning model at https://github.com/THUDM/CogVideo.

2975. One Step Diffusion via Shortcut Models

链接：https://iclr.cc/virtual/2025/poster/29802 abstract： Diffusion models and flow matching models have enabled generating diverse and realistic images by learning to transfer noise to data. However, sampling from these models involves iterative denoising over many neural network passes, making generation slow and expensive. Previous approaches for speeding up sampling require complex training regimes, such as multiple training phases, multiple networks, or fragile scheduling. We introduce Shortcut Models, a family of generative models that use a single network and training phase to produce high-quality samples in a single or multiple sampling steps. Shortcut models condition the network not only on the current noise level but also on the desired step size, allowing the model to skip ahead in the generation process. Across a wide range of sampling step budgets, shortcut models consistently produce higher quality samples than previous approaches, such as consistency models and reflow. Compared to distillation, shortcut models reduce complexity to a single network and training phase and additionally allow varying step budgets at inference time.

2976. OGBench: Benchmarking Offline Goal-Conditioned RL

链接：https://iclr.cc/virtual/2025/poster/29950 abstract： Offline goal-conditioned reinforcement learning (GCRL) is a major problem in reinforcement learning (RL) because it provides a simple, unsupervised, and domain-agnostic way to acquire diverse behaviors and representations from unlabeled data without rewards. Despite the importance of this setting, we lack a standard benchmark that can systematically evaluate the capabilities of offline GCRL algorithms. In this work, we propose OGBench, a new, high-quality benchmark for algorithms research in offline goal-conditioned RL. OGBench consists of 8 types of environments, 85 datasets, and reference implementations of 6 representative offline GCRL algorithms. We have designed these challenging and realistic environments and datasets to directly probe different capabilities of algorithms, such as stitching, long-horizon reasoning, and the ability to handle high-dimensional inputs and stochasticity. While representative algorithms may rank similarly on prior benchmarks, our experiments reveal stark strengths and weaknesses in these different capabilities,

providing a strong foundation for building new algorithms. Project page: https://seohong.me/projects/ogbench

2977. Random Is All You Need: Random Noise Injection on Feature Statistics for Generalizable Deep Image Denoising

链接：https://iclr.cc/virtual/2025/poster/27690 abstract：

2978. Prioritized Generative Replay

链接：https://iclr.cc/virtual/2025/poster/30955 abstract：

2979. CURIE: Evaluating LLMs on Multitask Scientific Long-Context Understanding and Reasoning

链接：https://iclr.cc/virtual/2025/poster/28609 abstract：

2980. Multimodal Quantitative Language for Generative Recommendation

链接：https://iclr.cc/virtual/2025/poster/27932 abstract： Generative recommendation has emerged as a promising paradigm aiming at directly generating the identifiers of the target candidates.Most existing methods attempt to leverage prior knowledge embedded in Pre-trained Language Models (PLMs) to improve the recommendation performance. However, they often fail to accommodate the differences between the general linguistic knowledge of PLMs and the specific needs of recommendation systems. Moreover, they rarely consider the complementary knowledge between the multimodal information of items, which represents the multi-faceted preferences of users. To facilitate efficient recommendation knowledge transfer, we propose a novel approach called Multimodal Quantitative Language for Generative Recommendation (MQL4GRec). Our key idea is to transform items from different domains and modalities into a unified language, which can serve as a bridge for transferring recommendation knowledge. Specifically, we first introduce quantitative translators to convert the text and image content of items from various domains into a new and concise language, known as quantitative language, with all items sharing the same vocabulary. Then, we design a series of quantitative language generation tasks to enrich quantitative language with semantic information and prior knowledge. Finally, we achieve the transfer of recommendation knowledge from different domains and modalities to the recommendation task through pre-training and fine-tuning. We evaluate the effectiveness of MQL4GRec through extensive experiments and comparisons with existing methods, achieving improvements over the baseline by 11.18\%, 14.82\%, and 7.95\% on the NDCG metric across three different datasets, respectively.

2981. FreSh: Frequency Shifting for Accelerated Neural Representation Learning

链接：https://iclr.cc/virtual/2025/poster/27671 abstract： Implicit Neural Representations (INRs) have recently gained attention as a powerful approach for continuously representing signals such as images, videos, and 3D shapes using multilayer perceptrons (MLPs). However, MLPs are known to exhibit a low-frequency bias, limiting their ability to capture high-frequency details accurately. This limitation is typically addressed by incorporating high-frequency input embeddings or specialized activation layers. In this work, we demonstrate that these embeddings and activations are often configured with hyperparameters that perform well on average but are suboptimal for specific input signals under consideration, necessitating a costly grid search to identify optimal settings. Our key observation is that the initial frequency spectrum of an untrained model's output correlates strongly with the model's eventual performance on a given target signal. Leveraging this insight, we propose frequency shifting (or FreSh), a method that selects embedding hyperparameters to align the frequency spectrum of the model’s initial output with that of the target signal. We show that this simple initialization technique improves performance across various neural representation methods and tasks, achieving results comparable to extensive hyperparameter sweeps but with only marginal computational overhead compared to training a single model with default hyperparameters.

2982. SVD-LLM: Truncation-aware Singular Value Decomposition for Large Language Model Compression

链接：https://iclr.cc/virtual/2025/poster/30003 abstract： The advancements in Large Language Models (LLMs) have been hindered bytheir substantial sizes, which necessitates LLM compression methods for practicaldeployment. Singular Value Decomposition (SVD) offers a promising solution forLLM compression. However, state-of-the-art SVD-based LLM compression meth-ods have two key limitations: truncating smaller singular values may lead to highercompression loss, and the lack of update on the compressed weights after SVDtruncation. In this work, we propose SVD-LLM, a SVD-based post-training LLMcompression method that addresses the limitations of existing methods. SVD-LLMincorporates a truncation-aware data whitening technique to ensure a direct map-ping between singular values and compression loss. Moreover, SVD-LLM adoptsa parameter update with sequential low-rank approximation to compensate forthe accuracy degradation after SVD compression. We evaluate SVD-LLM on 10datasets and seven models from three different LLM families at three differentscales. Our results demonstrate the superiority of SVD-LLM over state-of-the-arts,especially at high model compression ratios.

2983. Precedence-Constrained Winter Value for Effective Graph Data Valuation

链接：https://iclr.cc/virtual/2025/poster/28048 abstract： Data valuation is essential for quantifying data's worth, aiding in assessing data quality and determining fair compensation. While existing data valuation methods have proven effective in evaluating the value of Euclidean data, they face limitations when applied to the increasingly popular graph-structured data. Particularly, graph data valuation introduces unique challenges, primarily stemming from the intricate dependencies among nodes and the exponential growth in value estimation costs. To address the challenging problem of graph data valuation, we put forth an innovative solution, Precedence-Constrained Winter (PC-Winter) Value, to account for the complex graph structure. Furthermore, we develop a variety of strategies to address the computational challenges and enable efficient approximation of PC-Winter. Extensive experiments demonstrate the effectiveness of PC-Winter across diverse datasets and tasks.

2984. Seeing Eye to AI: Human Alignment via Gaze-Based Response Rewards for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/27970 abstract： Advancements in Natural Language Processing (NLP), have led to the emergence of Large Language Models (LLMs) such as GPT, Llama, Claude, and Gemini, which excel across a range of tasks but require extensive fine-tuning to align their outputs with human expectations. A widely used method for achieving this alignment is Reinforcement Learning from Human Feedback (RLHF), which, despite its success, faces challenges in accurately modelling human preferences. In this paper, we introduce GazeReward, a novel framework that integrates implicit feedback -- and specifically eye-tracking (ET) data -- into the Reward Model (RM). In addition, we explore how ET-based features can provide insights into user preferences. Through ablation studies we test our framework with different integration methods, LLMs, and ET generator models, demonstrating that our approach significantly improves the accuracy of the RM on established human preference datasets. This work advances the ongoing discussion on optimizing AI alignment with human values, exploring the potential of cognitive data for shaping future NLP research.

2985. Uncertainty Modeling in Graph Neural Networks via Stochastic Differential Equations

链接：https://iclr.cc/virtual/2025/poster/29534 abstract： We propose a novel Stochastic Differential Equation (SDE) framework to address the problem of learning uncertainty-aware representations for graph-structured data. While Graph Neural Ordinary Differential Equations (GNODEs) have shown promise in learning node representations, they lack the ability to quantify uncertainty. To address this, we introduce Latent Graph Neural Stochastic Differential Equations (LGNSDE), which enhance GNODE by embedding randomness through a Bayesian prior-posterior mechanism for epistemic uncertainty and Brownian motion for aleatoric uncertainty. By leveraging the existence and uniqueness of solutions to graph-based SDEs, we prove that the variance of the latent space bounds the variance of model outputs, thereby providing theoretically sensible guarantees for the uncertainty estimates. Furthermore, we show mathematically that LGNSDEs are robust to small perturbations in the input, maintaining stability over time. Empirical results across several benchmarks demonstrate that our framework is competitive in out-of-distribution detection, robustness to noise perturbations, and active learning, underscoring the ability of LGNSDEs to quantify uncertainty reliably.

2986. LASER: A Neuro-Symbolic Framework for Learning Spatio-Temporal Scene Graphs with Weak Supervision

链接：https://iclr.cc/virtual/2025/poster/30236 abstract： Supervised approaches for learning spatio-temporal scene graphs (STSG) from video are greatly hindered due to their reliance on STSG-annotated videos, which are labor-intensive to construct at scale. Is it feasible to instead use readily available video captions as weak supervision? To address this question, we propose LASER, a neuro-symbolic framework to enable training STSG generators using only video captions. LASER employs large language models to first extract logical specifications with rich spatio-temporal semantic information from video captions. LASER then trains the underlying STSG generator to align the predicted STSG with the specification. The alignment algorithm overcomes the challenges of weak supervision by leveraging a differentiable symbolic reasoner and using a combination of contrastive, temporal, and semantics losses. The overall approach efficiently trains low-level perception models to extract a fine-grained STSG that conforms to the video caption. In doing so, it enables a novel methodology for learning STSGs without tedious annotations. We evaluate our method on three video datasets: OpenPVSG, 20BN, and MUGEN. Our approach demonstrates substantial improvements over fully-supervised baselines, achieving a unary predicate prediction accuracy of 27.78% (+12.65%) and a binary recall@5 of 0.42 (+0.22) on OpenPVSG. Additionally, LASER exceeds baselines by 7% on 20BN and 5.2% on MUGEN in terms of overall predicate prediction accuracy.

2987. Omni-MATH: A Universal Olympiad Level Mathematic Benchmark for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/27714 abstract： Recent advancements in large language models (LLMs) have led to significant breakthroughs in mathematical reasoning capabilities. However, existing benchmarks like GSM8K or MATH are now

being solved with high accuracy (e.g., OpenAI o1 achieves 94.8% on MATH dataset), indicating their inadequacy for truly challenging these models. To bridge this gap, we propose a comprehensive and challenging benchmark specifically designed to assess LLMs' mathematical reasoning at the Olympiad level. Unlike existing Olympiad-related benchmarks, our dataset focuses exclusively on mathematics and comprises a vast collection of 4428 competition-level problems with rigorous human annotation. These problems are meticulously categorized into over 33 sub-domains and span more than 10 distinct difficulty levels, enabling a holistic assessment of model performance in Olympiad-mathematical reasoning. Furthermore, we conducted an in-depth analysis based on this benchmark. Our experimental results show that even the most advanced models, OpenAI o1-mini and OpenAI o1-preview, struggle with highly challenging Olympiad-level problems, with 60.54% and 52.55% accuracy, highlighting significant challenges in Olympiad-level mathematical reasoning.

2988. Federated Continual Learning Goes Online: Uncertainty-Aware Memory Management for Vision Tasks and Beyond

链接：https://iclr.cc/virtual/2025/poster/28896 abstract： Given the ability to model more realistic and dynamic problems, Federated Continual Learning (FCL) has been increasingly investigated recently. A well-known problem encountered in this setting is the so-called catastrophic forgetting, for which the learning model is inclined to focus on more recent tasks while forgetting the previously learned knowledge. The majority of the current approaches in FCL propose generative-based solutions to solve said problem. However, this setting requires multiple training epochs over the data, implying an offline setting where datasets are stored locally and remain unchanged over time. Furthermore, the proposed solutions are tailored for vision tasks solely. To overcome these limitations, we propose a new approach to deal with different modalities in the online scenario where new data arrive in streams of mini-batches that can only be processed once. To solve catastrophic forgetting, we propose an uncertainty-aware memory-based approach. Specifically, we suggest using an estimator based on the Bregman Information (BI) to compute the model's variance at the sample level. Through measures of predictive uncertainty, we retrieve samples with specific characteristics, and – by retraining the model on such samples – we demonstrate the potential of this approach to reduce the forgetting effect in realistic settings while maintaining data confidentiality and competitive communication efficiency compared to state-of-the-art approaches.

2989. Diversity-Rewarded CFG Distillation

链接：https://iclr.cc/virtual/2025/poster/32061 abstract： Generative models are transforming creative domains such as music generation, with inference-time strategies like Classifier-Free Guidance (CFG) playing a crucial role. However, CFG doubles inference cost while limiting originality and diversity across generated contents. In this paper, we introduce diversity-rewarded CFG distillation, a novel finetuning procedure that distills the strengths of CFG while addressing its limitations. Our approach optimises two training objectives: (1) a distillation objective, encouraging the model alone (without CFG) to imitate the CFG-augmented predictions, and (2) an RL objective with a diversity reward, promoting the generation of diverse outputs for a given prompt. By finetuning, we learn model weights with the ability to generate high-quality and diverse outputs, without any inference overhead. This also unlocks the potential of weight-based model merging strategies: by interpolating between the weights of two models (the first focusing on quality, the second on diversity), we can control the quality-diversity trade-off at deployment time, and even further boost performance. We conduct extensive experiments on the MusicLM text-to-music generative model, where our approach surpasses CFG in terms of quality-diversity Pareto optimality. According to human evaluators, our finetuned-then-merged model generates samples with higher quality-diversity than the base model augmented with CFG. Explore our generations at https://musicdiversity.github.io/.

2990. Gaussian Differentially Private Human Faces Under a Face Radial Curve Representation

链接：https://iclr.cc/virtual/2025/poster/30080 abstract： In this paper we consider the problem of releasing a Gaussian Differentially Private (GDP) 3D human face. The human face is a complex structure with many features and inherently tied to one's identity. Protecting this data, in a formally private way, is important yet challenging given the dimensionality of the problem. We extend approximate DP techniques for functional data to the GDP framework. We further propose a novel representation, face radial curves, of a 3D face as a set of functions and then utilize our proposed GDP functional data mechanism. To preserve the shape of the face while injecting noise we rely on tools from shape analysis for our novel representation of the face. We show that our method preserves the shape of the average face and injects less noise than traditional methods for the same privacy budget. Our mechanism consists of two primary components, the first is generally applicable to function value summaries (as are commonly found in nonparametric statistics or functional data analysis) while the second is general to disk-like surfaces and hence more applicable than just to human faces.

2991. MANTRA: The Manifold Triangulations Assemblage

链接：https://iclr.cc/virtual/2025/poster/29323 abstract： The rising interest in leveraging higher-order interactions present in complex systems hasled to a surge in more expressive models exploiting higher-order structures in the data,especially in topological deep learning (TDL), which designs neural networks on higher-order domains such as simplicial complexes. However, progress in this field is hinderedby the scarcity of datasets for benchmarking these architectures. To address this gap, weintroduce MANTRA, the first large-scale, diverse, and intrinsically higher-order dataset forbenchmarking higher-order models, comprising over 43,000 and 250,000 triangulationsof surfaces and three-dimensional manifolds, respectively. With MANTRA,

we assessseveral graph- and simplicial complex-based models on three topological classificationtasks. We demonstrate that while simplicial complex-based neural networks generallyoutperform their graph-based counterparts in capturing simple topological invariants, theyalso struggle, suggesting a rethink of TDL. Thus, MANTRA serves as a benchmark forassessing and advancing topological methods, paving the way towards more effectivehigher-order models.

2992. Diffusion-Based Planning for Autonomous Driving with Flexible Guidance

链接：https://iclr.cc/virtual/2025/poster/27852 abstract： Achieving human-like driving behaviors in complex open-world environments is a critical challenge in autonomous driving. Contemporary learning-based planning approaches such as imitation learning methods often struggle to balance competing objectives and lack of safety assurance,due to limited adaptability and inadequacy in learning complex multi-modal behaviors commonly exhibited in human planning, not to mention their strong reliance on the fallback strategy with predefined rules. We propose a novel transformer-based Diffusion Planner for closed-loop planning, which can effectively model multi-modal driving behavior and ensure trajectory quality without any rule-based refinement. Our model supports joint modeling of both prediction and planning tasks under the same architecture, enabling cooperative behaviors between vehicles. Moreover, by learning the gradient of the trajectory score function and employing a flexible classifier guidance mechanism, Diffusion Planner effectively achieves safe and adaptable planning behaviors. Evaluations on the large-scale real-world autonomous planning benchmark nuPlan and our newly collected 200-hour delivery-vehicle driving dataset demonstrate that Diffusion Planner achieves state-of-the-art closed-loop performance with robust transferability in diverse driving styles.

2993. Skill Expansion and Composition in Parameter Space

链接：https://iclr.cc/virtual/2025/poster/30286 abstract： Humans excel at reusing prior knowledge to address new challenges and developing skills while solving problems. This paradigm becomes increasingly popular in the development of autonomous agents, as it develops systems that can self-evolve in response to new challenges like human beings. However, previous methods suffer from limited training efficiency when expanding new skills and fail to fully leverage prior knowledge to facilitate new task learning. We propose Parametric Skill Expansion and Composition (PSEC), a new framework designed to iteratively evolve the agents' capabilities and efficiently address new challenges by maintaining a manageable skill library. This library can progressively integrate skill primitives as plug-and-play Low-Rank Adaptation (LoRA) modules in parameter-efficient finetuning, facilitating efficient and flexible skill expansion. This structure also enables the direct skill compositions in parameter space by merging LoRA modules that encode different skills, leveraging shared information across skills to effectively program new skills. Based on this, we propose a context-aware modular to dynamically activate different skills to collaboratively handle new tasks. Empowering diverse applications including multi-objective composition, dynamics shift, and continual policy shift, the results on D4RL, DSRL benchmarks, and the DeepMind Control Suite show that PSEC exhibits superior capacity to leverage prior knowledge to efficiently tackle new challenges, as well as expand its skill libraries to evolve the capabilities. Project website: https://ltlhuuu.github.io/PSEC/.

2994. What Do You See in Common? Learning Hierarchical Prototypes over Tree-of-Life to Discover Evolutionary Traits

链接：https://iclr.cc/virtual/2025/poster/30990 abstract： A grand challenge in biology is to discover evolutionary traits---features of organisms common to a group of species with a shared ancestor in the tree of life (also referred to as phylogenetic tree). With the growing availability of image repositories in biology, there is a tremendous opportunity to discover evolutionary traits directly from images in the form of a hierarchy of prototypes. However, current prototype-based methods are mostly designed to operate over a flat structure of classes and face several challenges in discovering hierarchical prototypes, including the issue of learning over-specific prototypes at internal nodes. To overcome these challenges, we introduce the framework of Hierarchy aligned Commonality through Prototypical Networks (HComP-Net). The key novelties in HComP-Net include a novel over-specificity loss to avoid learning over-specific prototypes, a novel discriminative loss to ensure prototypes at an internal node are absent in the contrasting set of species with different ancestry, and a novel masking module to allow for the exclusion of over-specific prototypes at higher levels of the tree without hampering classification performance. We empirically show that HComP-Net learns prototypes that are accurate, semantically consistent, and generalizable to unseen species in comparison to baselines. Our code is publicly accessible at Imageomics Institute Github site: https://github.com/Imageomics/HComPNet.

2995. Backdooring Vision-Language Models with Out-Of-Distribution Data

链接：https://iclr.cc/virtual/2025/poster/28044 abstract： The emergence of Vision-Language Models (VLMs) represents a significant advancement in integrating computer vision with Large Language Models (LLMs) to generate detailed text descriptions from visual inputs. Despite their growing importance, the security of VLMs, particularly against backdoor attacks, is under explored. Moreover, prior works often assume attackers have access to the original training data, which is often unrealistic. In this paper, we address a more practical and challenging scenario where attackers must rely solely on Out-Of-Distribution (OOD) data. We introduce VLOOD (Backdoor Vision-Language Models using Out-of-Distribution Data), a novel approach with two key contributions: (1) demonstrating backdoor attacks on VLMs in complex image-to-text tasks while minimizing degradation of the original semantics under poisoned inputs, and (2) proposing innovative techniques for backdoor injection without requiring any access to the original training data. Our evaluation on image captioning and visual question

answering (VQA) tasks confirms the effectiveness of VLOOD, revealing a critical security vulnerability in VLMs and laying the foundation for future research on securing multimodal models against sophisticated threats.

2996. Geometry of Long-Tailed Representation Learning: Rebalancing Features for Skewed Distributions

链接：https://iclr.cc/virtual/2025/poster/30248 abstract： Deep learning has achieved significant success by training on balanced datasets. However, real-world data often exhibit long-tailed distributions. Empirical studies have revealed that long-tailed data skew data representations, where head classes dominate the feature space. Many methods have been proposed to empirically rectify the skewed representations. However, a clear understanding of the underlying cause and extent of this skew remains lacking. In this study, we provide a comprehensive theoretical analysis to elucidate how long-tailed data affect feature distributions, deriving the conditions under which centers of tail classes shrink together or even collapse into a single point. This results in overlapping feature distributions of tail classes, making features in the overlapping regions inseparable. Moreover, we demonstrate that merely empirically correcting the skewed representations of the training data is insufficient to separate the overlapping features due to distribution shifts between the training and real data. To address these challenges, we propose a novel long-tailed representation learning method, FeatRecon. It reconstructs the feature space in order to arrange features from different classes into symmetricial and linearly separable regions. This, in turn, enhances the model's robustness to long-tailed data. We validate the effectiveness of our method through extensive experiments on the CIFAR-10-LT, CIFAR-100-LT, ImageNet-LT, and iNaturalist 2018 datasets.

2997. Gaussian-Det: Learning Closed-Surface Gaussians for 3D Object Detection

链接：https://iclr.cc/virtual/2025/poster/30438 abstract： Skins wrapping around our bodies, leathers covering over the sofa, sheet metal coating the car – it suggests that objects are enclosed by a series of continuous surfaces, which provides us with informative geometry prior for objectness deduction. In this paper, we propose Gaussian-Det which leverages Gaussian Splatting as surface representation for multi-view based 3D object detection. Unlike existing monocular or NeRF-based methods which depict the objects via discrete positional data, Gaussian-Det models the objects in a continuous manner by formulating the input Gaussians as feature descriptors on a mass of partial surfaces. Furthermore, to address the numerous outliers inherently introduced by Gaussian splatting, we accordingly devise a Closure Inferring Module (CIM) for the comprehensive surface-based objectness deduction. CIM firstly estimates the probabilistic feature residuals for partial surfaces given the underdetermined nature of Gaussian Splatting, which are then coalesced into a holistic representation on the overall surface closure of the object proposal. In this way, the surface information Gaussian-Det exploits serves as the prior on the quality and reliability of objectness and the information basis of proposal refinement. Experiments on both synthetic and real-world datasets demonstrate that Gaussian-Det outperforms various existing approaches, in terms of both average precision and recall.

2998. A Transfer Attack to Image Watermarks

链接：https://iclr.cc/virtual/2025/poster/29465 abstract： Watermark has been widely deployed by industry to detect AI-generated images. The robustness of such watermark-based detector against evasion attacks in the white-box and black-box settings is well understood in the literature. However, the robustness in the no-box setting is much less understood. In this work, we propose a new transfer evasion attack to image watermark in the no-box setting. Our transfer attack adds a perturbation to a watermarked image to evade multiple surrogate watermarking models trained by the attacker itself, and the perturbed watermarked image also evades the target watermarking model. Our major contribution is to show that, both theoretically and empirically, watermark-based AI-generated image detector based on existing watermarking methods is not robust to evasion attacks even if the attacker does not have access to the watermarking model nor the detection API. Our code is available at: https://github.com/hifi-hyp/Watermark-Transfer-Attack.

2999. Dysca: A Dynamic and Scalable Benchmark for Evaluating Perception Ability of LVLMs

链接：https://iclr.cc/virtual/2025/poster/29107 abstract： Currently many benchmarks have been proposed to evaluate the perception ability of the Large Vision-Language Models (LVLMs).However, most benchmarks conduct questions by selecting images from existing datasets, resulting in the potential data leakage. Besides, these benchmarks merely focus on evaluating LVLMs on the realistic style images and clean scenarios, leaving the multi-stylized images and noisy scenarios unexplored. In response to these challenges, we propose a dynamic and scalable benchmark named Dysca for evaluating LVLMs by leveraging synthesis images. Specifically, we leverage Stable Diffusion and design a rule-based method to dynamically generate novel images, questions and the corresponding answers. We consider 51 kinds of image styles and evaluate the perception capability in 20 subtasks. Moreover, we conduct evaluations under 4 scenarios (i.e., Clean, Corruption, Print Attacking and Adversarial Attacking) and 3 question types (i.e., Multi-choices, True-or-false and Free-form). Thanks to the generative paradigm, Dysca serves as a scalable benchmark for easily adding new subtasks and scenarios. A total of 24 advanced open-source LVLMs and 2 close-source LVLMs are evaluated on Dysca, revealing the drawbacks of current LVLMs. The benchmark is released in anonymous github page \url{https://github.com/Benchmark-Dysca/Dysca}.

3000. MM-EMBED: UNIVERSAL MULTIMODAL RETRIEVAL WITH MULTIMODAL LLMS

链接：https://iclr.cc/virtual/2025/poster/28720 abstract：