3001. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models

链接：https://iclr.cc/virtual/2025/poster/28505 abstract：

3002. GenXD: Generating Any 3D and 4D Scenes

链接：https://iclr.cc/virtual/2025/poster/31199 abstract： Recent developments in 2D visual generation have been remarkably successful. However, 3D and 4D generation remain challenging in real-world applications due to the lack of large-scale 4D data and effective model design. In this paper, we propose to jointly investigate general 3D and 4D generation by leveraging camera and object movements commonly observed in daily life. Due to the lack of real-world 4D data in the community, we first propose a data curation pipeline to obtain camera poses and object motion strength from videos. Based on this pipeline, we introduce a large-scale real-world 4D scene dataset: CamVid-30K. By leveraging all the 3D and 4D data, we develop our framework, GenXD, which allows us to produce any 3D or 4D scene. We propose multiview-temporal modules, which disentangle camera and object movements, to seamlessly learn from both 3D and 4D data. Additionally, GenXD employs masked latent conditions to support a variety of conditioning views. GenXD can generate videos that follow the camera trajectory as well as consistent 3D views that can be lifted into 3D representations. We perform extensive evaluations across various real-world and synthetic datasets, demonstrating GenXD's effectiveness and versatility compared to previous methods in 3D and 4D generation.

3003. Locality-aware Gaussian Compression for Fast and High-quality Rendering

链接：https://iclr.cc/virtual/2025/poster/28993 abstract： We present LocoGS, a locality-aware 3D Gaussian Splatting (3DGS) framework that exploits the spatial coherence of 3D Gaussians for compact modeling of volumetric scenes.To this end, we first analyze the local coherence of 3D Gaussian attributes, and propose a novel locality-aware 3D Gaussian representation that effectively encodes locally-coherent Gaussian attributes using a neural field representation with a minimal storage requirement.On top of the novel representation, LocoGS is carefully designed with additional components such as dense initialization, an adaptive spherical harmonics bandwidth scheme and different encoding schemes for different Gaussian attributes to maximize compression performance.Experimental results demonstrate that our approach outperforms the rendering quality of existing compact Gaussian representations for representative real-world 3D datasets while achieving from 54.6$\times$ to 96.6$\times$ compressed storage size and from 2.1$\times$ to 2.4$\times$ rendering speed than 3DGS. Even our approach also demonstrates an averaged 2.4$\times$ higher rendering speed than the state-of-the-art compression method with comparable compression performance.

3004. EditRoom: LLM-parameterized Graph Diffusion for Composable 3D Room Layout Editing

链接：https://iclr.cc/virtual/2025/poster/29277 abstract： Given the steep learning curve of professional 3D software and the time-consuming process of managing large 3D assets, language-guided 3D scene editing has significant potential in fields such as virtual reality, augmented reality, andgaming. However, recent approaches to language-guided 3D scene editing eitherrequire manual interventions or focus only on appearance modifications withoutsupporting comprehensive scene layout changes. In response, we propose EditRoom, a unified framework capable of executing a variety of layout edits throughnatural language commands, without requiring manual intervention. Specifically,EditRoom leverages Large Language Models (LLMs) for command planning andgenerates target scenes using a diffusion-based method, enabling six types of edits: rotate, translate, scale, replace, add, and remove. To addressthe lack of data for language-guided 3D scene editing, we have developed an automatic pipeline to augment existing 3D scene synthesis datasets and introducedEditRoom-DB, a large-scale dataset with 83k editing pairs, for training and evaluation. Our experiments demonstrate that our approach consistently outperformsother baselines across all metrics, indicating higher accuracy and coherence inlanguage-guided scene layout editing.

3005. MMWorld: Towards Multi-discipline Multi-faceted World Model Evaluation in Videos

链接：https://iclr.cc/virtual/2025/poster/28053 abstract： Multimodal Language Language Models (MLLMs) demonstrate the emerging abilities of "world models"---interpreting and reasoning about complex real-world dynamics. To assess these abilities, we posit videos are the ideal medium, as they encapsulate rich representations of real-world dynamics and causalities. To this end, we introduce MMWorld, a new benchmark for multi-discipline, multi-faceted multimodal video understanding. MMWorld distinguishes itself from previous video understanding benchmarks with two unique advantages: (1) multi-discipline, covering various disciplines that often require domain expertise for comprehensive understanding; (2) multi-faceted reasoning, including explanation, counterfactual thinking, future prediction, etc. MMWorld consists of a human-annotated dataset to evaluate MLLMs with questions about the whole videos and a synthetic dataset to analyze MLLMs within a single modality of perception. Together, MMWorld encompasses 1,910 videos across seven broad disciplines and 69 subdisciplines, complete with 6,627 question-answer pairs and associated captions. The evaluation includes 4 proprietary and 11 open-source MLLMs, which struggle on MMWorld (e.g., GPT-4o performs the best with only 62.5% accuracy), showing large room for improvement. Further

ablation studies reveal other interesting findings such as models' different skill sets from humans. We hope MMWorld can serve as an essential step towards world model evaluation in videos.

3006. SlowFast-VGen: Slow-Fast Learning for Action-Driven Long Video Generation

链接：https://iclr.cc/virtual/2025/poster/29480 abstract： Human beings are endowed with a complementary learning system, which bridges the slow learning of general world dynamics with fast storage of episodic memory from a new experience. Previous video generation models, however, primarily focus on slow learning by pre-training on vast amounts of data, overlooking the fast learning phase crucial for episodic memory storage. This oversight leads to inconsistencies across temporally distant frames when generating longer videos, as these frames fall beyond the model's context window. To this end, we introduce SlowFast-VGen, a novel dual-speed learning system for action-driven long video generation. Our approach incorporates a masked conditional video diffusion model for the slow learning of world dynamics, alongside an inference-time fast learning strategy based on a temporal LoRA module. Specifically, the fast learning process updates its temporal LoRA parameters based on local inputs and outputs, thereby efficiently storing episodic memory in its parameters. We further propose a slow-fast learning loop algorithm that seamlessly integrates the inner fast learning loop into the outer slow learning loop, enabling the recall of prior multi-episode experiences for context-aware skill learning. To facilitate the slow learning of an approximate world model, we collect a large-scale dataset of 200k videos with language action annotations, covering a wide range of scenarios. Extensive experiments show that SlowFast-VGen outperforms baselines across various metrics for action-driven video generation, achieving an FVD score of 514 compared to 782, and maintaining consistency in longer videos, with an average of 0.37 scene cuts versus 0.89. The slow-fast learning loop algorithm significantly enhances performances on long-horizon planning tasks as well.

3007. Meta-Continual Learning of Neural Fields

链接：https://iclr.cc/virtual/2025/poster/29837 abstract： Neural Fields (NF) have gained prominence as a versatile framework for complex data representation. This work unveils a new problem setting termed Meta-Continual Learning of Neural Fields (MCL-NF) and introduces a novel strategy that employs a modular architecture combined with optimization-based meta-learning. Focused on overcoming the limitations of existing methods for continual learning of neural fields, such as catastrophic forgetting and slow convergence, our strategy achieves high-quality reconstruction with significantly improved learning speed. We further introduce Fisher Information Maximization loss for neural radiance fields (FIM-NeRF), which maximizes information gains at the sample level to enhance learning generalization, with proved convergence guarantee and generalization bound. We perform extensive evaluations across image, audio, video reconstruction, and view synthesis tasks on six diverse datasets, demonstrating our method’s superiority in reconstruction quality and speed over existing MCL and CL-NF approaches. Notably, our approach attains rapid adaptation of neural fields for city-scale NeRF rendering with reduced parameter requirement.

3008. Faster Algorithms for Structured Linear and Kernel Support Vector Machines

链接：https://iclr.cc/virtual/2025/poster/30470 abstract： Quadratic programming is a ubiquitous prototype in convex programming. Many machine learning problems can be formulated as quadratic programming, including the famous Support Vector Machines (SVMs). Linear and kernel SVMs have been among the most popular models in machine learning over the past three decades, prior to the deep learning era.Generally, a quadratic program has an input size of $\Theta(n^2)$, where n is the number of variables. Assuming the Strong Exponential Time Hypothesis ($\textsf{SETH}$), it is known that no $O(n^{2-o(1)})$ time algorithm exists when the quadratic objective matrix is positive semidefinite (Backurs, Indyk, and Schmidt, NeurIPS'17). However, problems such as SVMs usually admit much smaller input sizes: one is given n data points, each of dimension d, and d is oftentimes much smaller than n. Furthermore, the SVM program has only $O(1)$ equality linear constraints. This suggests that faster algorithms are feasible, provided the program exhibits certain structures.In this work, we design the first nearly-linear time algorithm for solving quadratic programs whenever the quadratic objective admits a low-rank factorization, and the number of linear constraints is small. Consequently, we obtain results for SVMs:* For linear SVM when the input data is d-dimensional, our algorithm runs in time $\widetilde O(nd^{(\omega+1)/2}\log(1/\epsilon))$ where $\omega\approx 2.37$ is the fast matrix multiplication exponent; * For Gaussian kernel SVM, when the data dimension $d = O(\log n)$ and the squared dataset radius is sub-logarithmic in n, our algorithm runs in time $O(n^{1+o(1)}\log(1/\epsilon))$. We also prove that when the squared dataset radius is at least $\Omega(\log^2 n)$, then $\Omega(n^{2-o(1)})$ time is required. This improves upon the prior best lower bound in both the dimension d and the squared dataset radius.

3009. On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback

链接：https://iclr.cc/virtual/2025/poster/29353 abstract： As LLMs become more widely deployed, there is increasing interest in directly optimizing for feedback from end users (e.g. thumbs up) in addition to feedback from paid annotators. However, training to maximize human feedback creates a perverse incentive structure for the AI to resort to manipulative or deceptive tactics to obtain positive feedback from users who are vulnerable to such strategies. We study this phenomenon by training LLMs with Reinforcement Learning with simulated user feedback in environments of practical LLM usage. In our settings, we find that: 1)

Extreme forms of "feedback gaming" such as manipulation and deception are learned reliably; 2) Even if only 2% of users are vulnerable to manipulative strategies, LLMs learn to identify and target them while behaving appropriately with other users, making such behaviors harder to detect; 3) To mitigate this issue, it may seem promising to leverage continued safety training or LLM-as-judges during training to filter problematic outputs. Instead, we found that while such approaches help in some of our settings, they backfire in others, sometimes even leading to subtler manipulative behaviors. We hope our results can serve as a case study which highlights the risks of using gameable feedback sources -- such as user feedback -- as a target for RL. Our code is publicly available. Warning: some of our examples may be upsetting.

3010. Semi-Supervised Vision-Centric 3D Occupancy World Model for Autonomous Driving

链接：https://iclr.cc/virtual/2025/poster/28213 abstract：

3011. Adversarial Attacks on Data Attribution

链接：https://iclr.cc/virtual/2025/poster/28363 abstract： Data attribution aims to quantify the contribution of individual training data points to the outputs of an AI model, which has been used to measure the value of training data and compensate data providers. Given the impact on financial decisions and compensation mechanisms, a critical question arises concerning the adversarial robustness of data attribution methods. However, there has been little to no systematic research addressing this issue. In this work, we aim to bridge this gap by detailing a threat model with clear assumptions about the adversary's goal and capabilities and proposing principled adversarial attack methods on data attribution. We present two methods, Shadow Attack and Outlier Attack, which generate manipulated datasets to inflate the compensation adversarially. The Shadow Attack leverages knowledge about the data distribution in the AI applications, and derives adversarial perturbations through "shadow training", a technique commonly used in membership inference attacks. In contrast, the Outlier Attack does not assume any knowledge about the data distribution and relies solely on black-box queries to the target model's predictions. It exploits an inductive bias present in many data attribution methods - outlier data points are more likely to be influential - and employs adversarial examples to generate manipulated datasets. Empirically, in image classification and text generation tasks, the Shadow Attack can inflate the data-attribution-based compensation by at least 200%, while the Outlier Attack achieves compensation inflation ranging from 185% to as much as 643%. Our implementation is ready at https://github.com/TRAIS-Lab/adversarial-attack-data-attribution

3012. Energy-Weighted Flow Matching for Offline Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/30241 abstract： This paper investigates energy guidance in generative modeling, where the target distribution is defined as $q(\mathbf x) \propto p(\mathbf x)\exp(-\beta \mathcal E(\mathbf x))$, with $p(\mathbf x)$ being the data distribution and $\mathcal E(\mathbf x)$ as the energy function. To comply with energy guidance, existing methods often require auxiliary procedures to learn intermediate guidance during the diffusion process. To overcome this limitation, we explore energy-guided flow matching, a generalized form of the diffusion process. We introduce energy-weighted flow matching (EFM), a method that directly learns the energy-guided flow without the need for auxiliary models. Theoretical analysis shows that energy-weighted flow matching accurately captures the guided flow. Additionally, we extend this methodology to energy-weighted diffusion models and apply it to offline reinforcement learning (RL) by proposing the Q-weighted Iterative Policy Optimization (QIPO). Empirically, we demonstrate that the proposed QIPO algorithm improves performance in offline RL tasks. Notably, our algorithm is the first energy-guided diffusion model that operates independently of auxiliary models and the first exact energy-guided flow matching model in the literature.

3013. DPLM-2: A Multimodal Diffusion Protein Language Model

链接：https://iclr.cc/virtual/2025/poster/30917 abstract： Proteins are essential macromolecules defined by their amino acid sequences, which determine their three-dimensional structures and, consequently, their functions in all living organisms. Therefore, generative protein modeling necessitates a multimodal approach to simultaneously model, understand, and generate both sequences and structures. However, existing methods typically use separate models for each modality, limiting their ability to capture the intricate relationships between sequence and structure. This results in suboptimal performance in tasks that requires joint understanding and generation of both modalities.In this paper, we introduce DPLM-2, a multimodal protein foundation model that extends discrete diffusion protein language model (DPLM) to accommodate both sequences and structures.To enable structural learning with the language model, 3D coordinates are converted to discrete tokens using a lookup-free quantization-based tokenizer.By training on both experimental and high-quality synthetic structures, DPLM-2 learns the joint distribution of sequence and structure, as well as their marginals and conditionals.We also implement an efficient warm-up strategy to exploit the connection between large-scale evolutionary data and structural inductive biases from pre-trained sequence-based protein language models.Empirical evaluation shows that DPLM-2 can simultaneously generate highly compatible amino acid sequences and their corresponding 3D structures eliminating the need for a two-stage generation approach.Moreover, DPLM-2 demonstrates competitive performance in various conditional generation tasks, including folding, inverse folding, and scaffolding with multimodal motif inputs.

3014. CONTRA: Conformal Prediction Region via Normalizing Flow

Transformation

链接：https://iclr.cc/virtual/2025/poster/28304 abstract： Density estimation and reliable prediction regions for outputs are crucial in supervised and unsupervised learning. While conformal prediction effectively generates coverage-guaranteed regions, it struggles with multi-dimensional outputs due to reliance on one-dimensional nonconformity scores. To address this, we introduce CONTRA: CONformal prediction region via normalizing flow TRAnsformation. CONTRA utilizes the latent spaces of normalizing flows to define nonconformity scores based on distances from the center. This allows for the mapping of high-density regions in latent space to sharp prediction regions in the output space, surpassing traditional hyperrectangular or elliptical conformal regions. Further, for scenarios where other predictive models are favored over flow-based models, we extend CONTRA to enhance any such model with a reliable prediction region by training a simple normalizing flow on the residuals. We demonstrate that both CONTRA and its extension maintain guaranteed coverage probability and outperform existing methods in generating accurate prediction regions across various datasets. We conclude that CONTRA is an effective tool for (conditional) density estimation, addressing the under-explored challenge of delivering multi-dimensional prediction regions.

3015. OpenVid-1M: A Large-Scale High-Quality Dataset for Text-to-video Generation

链接：https://iclr.cc/virtual/2025/poster/28653 abstract： Text-to-video (T2V) generation has recently garnered significant attention thanks to the large multi-modality model Sora. However, T2V generation still faces two important challenges: 1) Lacking a precise open sourced high-quality dataset. The previously popular video datasets, e.g.WebVid-10M and Panda-70M, overly emphasized large scale, resulting in the inclusion of many low-quality videos andshort, imprecise captions. Therefore, it is challenging but crucial to collect a precise high-quality dataset while maintaining a scale of millions for T2V generation. 2) Ignoring to fully utilize textual information. Recent T2V methods have focused on vision transformers, using a simple cross attention module for video generation, which falls short of making full use of semantic information from text tokens. To address these issues, we introduce OpenVid-1M, a precise high-quality dataset with expressive captions. This open-scenario dataset contains over 1 million text-video pairs, facilitating research on T2V generation. Furthermore, we curate 433K 1080p videos from OpenVid-1M to create OpenVidHD-0.4M, advancing high-definition video generation. Additionally, we propose a novel Multi-modal Video Diffusion Transformer (MVDiT) capable of mining both structure information from visual tokens and semantic information from text tokens. Extensive experiments and ablation studies verify the superiority of OpenVid-1M over previous datasets and the effectiveness of our MVDiT.

3016. Show-o: One Single Transformer to Unify Multimodal Understanding and Generation

链接：https://iclr.cc/virtual/2025/poster/28376 abstract： We present a unified transformer, i.e., Show-o, that unifies multimodal understanding and generation. Unlike fully autoregressive models, Show-o unifies autoregressive and (discrete) diffusion modeling to adaptively handle inputs and outputs of various and mixed modalities. The unified model flexibly supports a wide range of vision-language tasks including visual question-answering, text-to-image generation, text-guided inpainting/extrapolation, and mixed-modality generation. Across various benchmarks, it demonstrates comparable or superior performance to existing individual models with an equivalent or larger number of parameters tailored for understanding or generation. This significantly highlights its potential as a next-generation foundation model.

3017. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models

链接：https://iclr.cc/virtual/2025/poster/31486 abstract： Large language models (LLMs) specializing in natural language generation (NLG) have recently started exhibiting promising capabilities across a variety of domains. However, gauging the trustworthiness of responses generated by LLMs remains an open challenge, with limited research on uncertainty quantification (UQ) for NLG. Furthermore, existing literature typically assumes white-box access to language models, which is becoming unrealistic either due to the closed-source nature of the latest LLMs or computational constraints. In this work, we investigate UQ in NLG for black-box LLMs. We first differentiate uncertainty vs confidence: the former refers to the ``dispersion'' of the potential predictions for a fixed input, and the latter refers to the confidence on a particular prediction/generation. We then propose and compare several confidence/uncertainty measures, applying them to selective NLG where unreliable results could either be ignored or yielded for further assessment. Experiments were carried out with several popular LLMs on question-answering datasets (for evaluation purposes). Results reveal that a simple measure for the semantic dispersion can be a reliable predictor of the quality of LLM responses, providing valuable insights for practitioners on uncertainty management when adopting LLMs.

3018. UTILITY: Utilizing Explainable Reinforcement Learning to Improve Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29522 abstract： Reinforcement learning (RL) faces two challenges: (1) The RL agent lacks explainability. (2) The trained RL agent is, in many cases, non-optimal and even far from optimal. To address the first challenge, explainable reinforcement learning (XRL) is proposed to explain the decision-making of the RL agent. In this paper,

we demonstrate that XRL can also be used to address the second challenge, i.e., improve RL performance. Our method has two parts. The first part provides a two-level explanation for why the RL agent is not optimal by identifying the mistakes made by the RL agent. Since this explanation includes the mistakes of the RL agent, it has the potential to help correct the mistakes and thus improve RL performance. The second part formulates a constrained bi-level optimization problem to learn how to best utilize the two-level explanation to improve RL performance. In specific, the upper level learns how to use the high-level explanation to shape the reward so that the corresponding policy can maximize the cumulative ground truth reward, and the lower level learns the corresponding policy by solving a constrained RL problem formulated using the low-level explanation. We propose a novel algorithm to solve this constrained bi-level optimization problem, and theoretically guarantee that the algorithm attains global optimality. We use MuJoCo experiments to show that our method outperforms state-of-the-art baselines.

3019. Efficient Diversity-Preserving Diffusion Alignment via Gradient-Informed GFlowNets

链接：https://iclr.cc/virtual/2025/poster/30600 abstract： While one commonly trains large diffusion models by collecting datasets on target downstream tasks, it is often desired to align and finetune pretrained diffusion models with some reward functions that are either designed by experts or learned from small-scale datasets. Existing post-training methods for reward finetuning of diffusion models typically suffer from lack of diversity in generated samples, lack of prior preservation, and/or slow convergence in finetuning. Inspired by recent successes in generative flow networks (GFlowNets), a class of probabilistic models that sample with the unnormalized density of a reward function, we propose a novel GFlowNet method dubbed Nabla-GFlowNet (abbreviated as \nabla-GFlowNet), the first GFlowNet method that leverages the rich signal in reward gradients, together with an objective called \nabla-DB plus its variant residual \nabla-DB designed for prior-preserving diffusion finetuning. We show that our proposed method achieves fast yet diversity- and prior-preserving finetuning of Stable Diffusion, a large-scale text-conditioned image diffusion model, on different realistic reward functions.

3020. Can Large Language Models Understand Symbolic Graphics Programs?

链接：https://iclr.cc/virtual/2025/poster/29250 abstract： Against the backdrop of enthusiasm for large language models (LLMs), there is a growing need to scientifically assess their capabilities and shortcomings. This is nontrivial in part because it is difficult to find tasks which the models have not encountered during training. Utilizing symbolic graphics programs, we propose a domain well-suited to test multiple spatial-semantic reasoning skills of LLMs. Popular in computer graphics, these programs procedurally generate visual data. While LLMs exhibit impressive skills in general program synthesis and analysis, symbolic graphics programs offer a new layer of evaluation: they allow us to test an LLM's ability to answer semantic questions about the images or 3D geometries without a vision encoder. To semantically understand the symbolic programs, LLMs would need to possess the ability to "imagine" and reason how the corresponding graphics content would look with only the symbolic description of the local curvatures and strokes. We use this task to evaluate LLMs by creating a large benchmark for the semantic visual understanding of symbolic graphics programs, built procedurally with minimal human effort. Particular emphasis is placed on transformations of images that leave the image level semantics invariant while introducing significant changes to the underlying program. We evaluate commercial and open-source LLMs on our benchmark to assess their ability to reason about visual output of programs, finding that LLMs considered stronger at reasoning generally perform better. Lastly, we introduce a novel method to improve this ability -- Symbolic Instruction Tuning (SIT), in which the LLM is finetuned with pre-collected instruction data on symbolic graphics programs. Interestingly, we find that SIT not only improves LLM's understanding on symbolic programs, but it also improves general reasoning ability on various other benchmarks.

3021. DistillHGNN: A Knowledge Distillation Approach for High-Speed Hypergraph Neural Networks

链接：https://iclr.cc/virtual/2025/poster/27878 abstract： In this paper, we propose a novel framework to significantly enhance the inference speed and memory efficiency of Hypergraph Neural Networks (HGNNs) while preserving their high accuracy. Our approach utilizes an advanced teacher-student knowledge distillation strategy. The teacher model, consisting of an HGNN and a Multi-Layer Perceptron (MLP), not only produces soft labels but also transfers structural and high-order information to a lightweight Graph Convolutional Network (GCN) known as TinyGCN. This dual transfer mechanism enables the student model to effectively capture complex dependencies while benefiting from the faster inference and lower computational cost of the lightweight GCN. The student model is trained using both labeled data and soft labels provided by the teacher, with contrastive learning further ensuring that the student retains high-order relationships. This makes the proposed method efficient and suitable for real-time applications, achieving performance comparable to traditional HGNNs but with significantly reduced resource requirements.

3022. Deep Kernel Relative Test for Machine-generated Text Detection

链接：https://iclr.cc/virtual/2025/poster/27688 abstract： Recent studies demonstrate that two-sample test can effectively detect machine-generated texts (MGTs) with excellent adaptation ability to texts generated by newer LLMs. However, two-sample test-based detection relies on the assumption that human-written texts (HWTs) must follow the distribution of seen HWTs. As a result, it tends to make mistakes in identifying HWTs that deviate from the seen HWT distribution, limiting their use in sensitive areas

like academic integrity verification. To address this issue, we propose to employ non-parametric kernel relative test to detect MGTs by testing whether it is statistically significant that the distribution of a text to be tested is closer to the distribution of HWTs than to the MGTs' distribution. We further develop a kernel optimisation algorithm in relative test to select the best kernel that can enhance the testing capability for MGT detection. As relative test does not assume that a text to be tested must belong exclusively to either MGTs or HWTs, relative test can largely reduce the false positive error compared to two-sample test, offering significant advantages in practice. Extensive experiments demonstrate the superior performance of our method, compared to state-of-the-art non-parametric and parametric detectors. The code and demo are available: https://github.com/xLearn-AU/R-Detect.

3023. Continuous Autoregressive Modeling with Stochastic Monotonic Alignment for Speech Synthesis

链接：https://iclr.cc/virtual/2025/poster/29023 abstract： We propose a novel autoregressive modeling approach for speech synthesis, combining a variational autoencoder (VAE) with a multi-modal latent space and an autoregressive model that uses Gaussian Mixture Models (GMM) as the conditional probability distribution. Unlike previous methods that rely on residual vector quantization, our model leverages continuous speech representations from the VAE's latent space, greatly simplifying the training and inference pipelines. We also introduce a stochastic monotonic alignment mechanism to enforce strict monotonic alignments. Our approach significantly outperforms the state-of-the-art autoregressive model VALL-E in both subjective and objective evaluations, achieving these results with only 10.3\% of VALL-E's parameters. This demonstrates the potential of continuous speech language models as a more efficient alternative to existing quantization-based speech language models. Sample audio can be found at \url{https://tinyurl.com/gmm-lm-tts}.

3024. Trained Transformer Classifiers Generalize and Exhibit Benign Overfitting In-Context

链接：https://iclr.cc/virtual/2025/poster/28608 abstract： Transformers have the capacity to act as supervised learning algorithms: by properly encoding a set of labeled training ("in-context") examples and an unlabeled test example into an input sequence of vectors of the same dimension, the forward pass of the transformer can produce predictions for that unlabeled test example. A line of recent work has shown that when linear transformers are pre-trained on random instances for linear regression tasks, these trained transformers make predictions using an algorithm similar to that of ordinary least squares. In this work, we investigate the behavior of linear transformers trained on random linear classification tasks. Via an analysis of the implicit regularization of gradient descent, we characterize how many pre-training tasks and in-context examples are needed for the trained transformer to generalize well at test-time. We further show that in some settings, these trained transformers can exhibit "benign overfitting in-context": when in-context examples are corrupted by label flipping noise, the transformer memorizes all of its in-context examples (including those with noisy labels) yet still generalizes near-optimally for clean test examples.

3025. Mutual Reasoning Makes Smaller LLMs Stronger Problem-Solver

链接：https://iclr.cc/virtual/2025/poster/30877 abstract： This paper introduces rStar, a self-play mutual reasoning approach that significantly improves reasoning capabilities of small language models (SLMs) without fine-tuning or superior models. rStar decouples reasoning into a self-play mutual generation-discrimination process. First, a target SLM augments the Monte Carlo Tree Search (MCTS) with a rich set of human-like reasoning actions to construct higher quality reasoning trajectories. Next, another SLM, with capabilities similar to the target SLM, acts as a discriminator to verify each trajectory generated by the target SLM. The mutually agreed reasoning trajectories are considered mutual consistent, thus are more likely to be correct. Extensive experiments across five SLMs demonstrate rStar can effectively solve diverse reasoning problems, including GSM8K, GSM-Hard, MATH, SVAMP, and StrategyQA. Remarkably, rStar boosts GSM8K accuracy from 12.51\% to 63.91\% for LLaMA2-7B, from 36.46\% to 81.88\% for Mistral-7B, from 74.53\% to 91.13\% for LLaMA3-8B-Instruct. Code is available at https://github.com/zhentingqi/rStar.

3026. Follow My Instruction and Spill the Beans: Scalable Data Extraction from Retrieval-Augmented Generation Systems

链接：https://iclr.cc/virtual/2025/poster/29275 abstract： Retrieval-Augmented Generation (RAG) improves pre-trained models by incorporating external knowledge at test time to enable customized adaptation. We study the risk of datastore leakage in Retrieval-In-Context RAG Language Models (LMs). We show that an adversary can exploit LMs' instruction-following capabilities to easily extract text data verbatim from the datastore of RAG systems built with instruction-tuned LMs via prompt injection. The vulnerability exists for a wide range of modern LMs that span Llama2, Mistral/Mixtral, Vicuna, SOLAR, WizardLM, Qwen1.5, and Platypus2, and the exploitability exacerbates as the model size scales up. We also study multiple effects of RAG setup on the extractability of data, indicating that following unexpected instructions to regurgitate data can be an outcome of failure in effectively utilizing contexts for modern LMs, and further show that such vulnerability can be greatly mitigated by position bias elimination strategies. Extending our study to production RAG models, GPTs, we design an attack that can cause datastore leakage with a near-perfect success rate on 25 randomly selected customized GPTs with at most 2 queries, and we extract text data verbatim at a rate of 41\% from a book of 77,000 words and 3\% from a corpus of 1,569,000 words by prompting the GPTs with only 100 queries generated by themselves.

3027. Visual Agents as Fast and Slow Thinkers

链接：https://iclr.cc/virtual/2025/poster/28402 abstract： Achieving human-level intelligence requires refining cognitive distinctions between \textit{System 1} and \textit{System 2} thinking. While contemporary AI, driven by large language models, demonstrates human-like traits, it falls short of genuine cognition. Transitioning from structured benchmarks to real-world scenarios presents challenges for visual agents, often leading to inaccurate and overly confident responses. To address the challenge, we introduce \textbf{\textsc{FaST}}, which incorporates the \textbf{Fa}st and \textbf{S}low \textbf{T}hinking mechanism into visual agents. \textsc{FaST} employs a switch adapter to dynamically select between \textit{System 1/2} modes, tailoring the problem-solving approach to different task complexity. It tackles uncertain and unseen objects by adjusting model confidence and integrating new contextual data. With this novel design, we advocate a \textit{flexible system}, \textit{hierarchical reasoning} capabilities, and a \textit{transparent decision-making} pipeline, all of which contribute to its ability to emulate human-like cognitive processes in visual intelligence. Empirical results demonstrate that \textsc{FaST} outperforms various well-known baselines, achieving 80.8\% accuracy over VQA^{v2} for visual question answering and 48.7\% $GIoU$ score over ReasonSeg for reasoning segmentation, demonstrate \textsc{FaST}'s superior performance. Extensive testing validates the efficacy and robustness of \textsc{FaST}'s core components, showcasing its potential to advance the development of cognitive visual agents in AI systems.

3028. Agent Security Bench (ASB): Formalizing and Benchmarking Attacks and Defenses in LLM-based Agents

链接：https://iclr.cc/virtual/2025/poster/29432 abstract： Although LLM-based agents, powered by Large Language Models (LLMs), can use external tools and memory mechanisms to solve complex real-world tasks, they may also introduce critical security vulnerabilities. However, the existing literature does not comprehensively evaluate attacks and defenses against LLM-based agents. To address this, we introduce Agent Security Bench (ASB), a comprehensive framework designed to formalize, benchmark, and evaluate the attacks and defenses of LLM-based agents, including 10 scenarios (e.g., e-commerce, autonomous driving, finance), 10 agents targeting the scenarios, over 400 tools, 27 different types of attack/defense methods, and 7 evaluation metrics. Based on ASB, we benchmark 10 prompt injection attacks, a memory poisoning attack, a novel Plan-of-Thought backdoor attack, 4 mixed attacks, and 11 corresponding defenses across 13 LLM backbones. Our benchmark results reveal critical vulnerabilities in different stages of agent operation, including system prompt, user prompt handling, tool usage, and memory retrieval, with the highest average attack success rate of 84.30\%, but limited effectiveness shown in current defenses, unveiling important works to be done in terms of agent security for the community. We also introduce a new metric to evaluate the agents' capability to balance utility and security. Our code can be found at https://github.com/agiresearch/ASB.

3029. Grounding Multimodal Large Language Model in GUI World

链接：https://iclr.cc/virtual/2025/poster/32092 abstract： Recent advancements in Multimodal Large Language Models (MLLMs) have accelerated the development of Graphical User Interface (GUI) agents capable of automating complex tasks across digital platforms. However, precise GUI element grounding remains a key challenge for accurate interaction and generalization. In this work, we present an effective GUI grounding framework, which includes an automated data collection engine that gathers extensive GUI screenshots and annotations to ensure broad generalization. We also propose a lightweight and flexible GUI grounding module designed to efficiently localize UI elements by pre-training on the collected data, and introduce a novel method to integrate this module with MLLMs for the effective execution of GUI tasks. Our approach demonstrates superior performance in task accuracy and adaptability, as validated by benchmarks such as ScreenSpot, MiniWob, AITW, and Mind2Web.

3030. Denoising with a Joint-Embedding Predictive Architecture

链接：https://iclr.cc/virtual/2025/poster/29012 abstract： Joint-embedding predictive architectures (JEPAs) have shown substantial promise in self-supervised representation learning, yet their application in generative modeling remains underexplored. Conversely, diffusion models have demonstrated significant efficacy in modeling arbitrary probability distributions. In this paper, we introduce Denoising with a Joint-Embedding Predictive Architecture (D-JEPA), pioneering the integration of JEPA within generative modeling. By recognizing JEPA as a form of masked image modeling, we reinterpret it as a generalized next-token prediction strategy, facilitating data generation in an auto-regressive manner. Furthermore, we incorporate diffusion loss to model the per-token probability distribution, enabling data generation in a continuous space. We also adapt flow matching loss as an alternative to diffusion loss, thereby enhancing the flexibility of D-JEPA. Empirically, with increased GFLOPs, D-JEPA consistently achieves lower FID scores with fewer training epochs, indicating its good scalability. Our base, large, and huge models outperform all previous generative models across all scales on ImageNet conditional generation benchmarks. Beyond image generation, D-JEPA is well-suited for other continuous data modeling, including video and audio.

3031. LoR-VP: Low-Rank Visual Prompting for Efficient Vision Model Adaptation

链接：https://iclr.cc/virtual/2025/poster/30935 abstract： Visual prompting has gained popularity as a method for adapting pre-

trained models to specific tasks, particularly in the realm of parameter-efficient tuning. However, existing visual prompting techniques often pad the prompt parameters around the image, limiting the interaction between the visual prompts and the original image to a small set of patches while neglecting the inductive bias present in shared information across different patches. In this study, we conduct a thorough preliminary investigation to identify and address these limitations. We propose a novel visual prompt design, introducing **Lo**w-**R**ank matrix multiplication for **V**isual **P**rompting (LoR-VP), which enables shared and patch-specific information across rows and columns of image pixels. Extensive experiments across seven network architectures and four datasets demonstrate significant improvements in both performance and efficiency compared to state-of-the-art visual prompting methods, achieving up to $6\times$ faster training times, utilizing $18\times$ fewer visual prompt parameters, and delivering a 3.1% improvement in performance.

3032. CofCA: A STEP-WISE Counterfactual Multi-hop QA benchmark

链接：https://iclr.cc/virtual/2025/poster/28269 abstract： While Large Language Models (LLMs) excel in question-answering (QA) tasks, their real reasoning abilities on multiple evidence retrieval and integration on Multi-hop QA tasks remain less explored. Firstly, LLMs sometimes generate answers that rely on internal memory rather than retrieving evidence and reasoning in the given context, which brings concerns about the evaluation quality of real reasoning abilities. Although previous counterfactual QA benchmarks can separate the internal memory of LLMs, they focus solely on final QA performance, which is insufficient for reporting LLMs' real reasoning abilities. Because LLMs are expected to engage in intricate reasoning processes that involve evidence retrieval and answering a series of sub-questions from given passages. Moreover, current factual Multi-hop QA (MHQA) benchmarks are annotated on open-source corpora such as Wikipedia, although useful for multi-step reasoning evaluation, they show limitations due to the potential data contamination in LLMs' pre-training stage. To address these issues, we introduce the Step-wise and Counterfactual benchmark (CofCA), a novel evaluation benchmark consisting of factual data and counterfactual data that reveals LLMs' real reasoning abilities on multi-step reasoning and reasoning chain evaluation. Our experimental results reveal a significant performance gap of several LLMs between Wikipedia-based factual data and counterfactual data, deeming data contamination issues in existing benchmarks. Moreover, we observe that LLMs usually bypass the correct reasoning chain, showing an inflated multi-step reasoning performance. We believe that our CofCA benchmark will enhance and facilitate the evaluations of trustworthy LLMs.

3033. Learning Molecular Representation in a Cell

链接：https://iclr.cc/virtual/2025/poster/30563 abstract： Predicting drug efficacy and safety in vivo requires information on biological responses (e.g., cell morphology and gene expression) to small molecule perturbations. However, current molecular representation learning methods do not provide a comprehensive view of cell states under these perturbations and struggle to remove noise, hindering model generalization. We introduce the Information Alignment (InfoAlign) approach to learn molecular representations through the information bottleneck method in cells. We integrate molecules and cellular response data as nodes into a context graph, connecting them with weighted edges based on chemical, biological, and computational criteria. For each molecule in a training batch, InfoAlign optimizes the encoder's latent representation with a minimality objective to discard redundant structural information. A sufficiency objective decodes the representation to align with different feature spaces from the molecule's neighborhood in the context graph. We demonstrate that the proposed sufficiency objective for alignment is tighter than existing encoder-based contrastive methods. Empirically, we validate representations from InfoAlign in two downstream applications: molecular property prediction against up to 27 baseline methods across four datasets, plus zero-shot molecule-morphology matching. The code and model are available at https://github.com/liugangcode/InfoAlign.

3034. MM1.5: Methods, Analysis & Insights from Multimodal LLM Fine-tuning

链接：https://iclr.cc/virtual/2025/poster/30222 abstract： We present MM1.5, a new family of multimodal large language models (MLLMs) designed to enhance capabilities in text-rich image understanding, visual referring and grounding, and multi-image reasoning. Building upon the MM1 architecture, MM1.5 adopts a data-centric approach to model training, systematically exploring the impact of diverse data mixtures across the entire model training lifecycle. This includes high-quality OCR data and synthetic captions for continual pre-training, as well as an optimized visual instruction-tuning data mixture for supervised fine-tuning. Our models range from 1B to 30B parameters, encompassing both dense and mixture-of-experts (MoE) variants, and demonstrate that careful data curation and training strategies can yield strong performance even at small scales (1B and 3B). Additionally, we introduce two specialized variants: MM1.5-Video, designed for video understanding, and MM1.5-UI, tailored for mobile UI understanding. Through extensive empirical studies and ablations, we provide detailed insights into the training processes and decisions that inform our final designs, offering valuable guidance for future research in MLLM development.

3035. Exact Certification of (Graph) Neural Networks Against Label Poisoning

链接：https://iclr.cc/virtual/2025/poster/29003 abstract： Machine learning models are highly vulnerable to label flipping, i.e., the adversarial modification (poisoning) of training labels to compromise performance. Thus, deriving robustness certificates is important to guarantee that test predictions remain unaffected and to understand worst-case robustness behavior. However, for Graph Neural Networks (GNNs), the problem of certifying label flipping has so far been unsolved. We change this by introducing an exact certification method, deriving both sample-wise and collective certificates. Our method leverages the Neural Tangent

Kernel (NTK) to capture the training dynamics of wide networks enabling us to reformulate the bilevel optimization problem representing label flipping into a Mixed-Integer Linear Program (MILP). We apply our method to certify a broad range of GNN architectures in node classification tasks. Thereby, concerning the worst-case robustness to label flipping: (i) we establish hierarchies of GNNs on different benchmark graphs; (ii) quantify the effect of architectural choices such as activations, depth and skip-connections; and surprisingly, (iii) uncover a novel phenomenon of the robustness plateauing for intermediate perturbation budgets across all investigated datasets and architectures. While we focus on GNNs, our certificates are applicable to sufficiently wide NNs in general through their NTK. Thus, our work presents the first exact certificate to a poisoning attack ever derived for neural networks, which could be of independent interest. The code is available at https://github.com/saper0/qpcert.

3036. Learning stochastic dynamics from snapshots through regularized unbalanced optimal transport

链接：https://iclr.cc/virtual/2025/poster/28822 abstract： Reconstructing dynamics using samples from sparsely time-resolved snapshots is an important problem in both natural sciences and machine learning. Here, we introduce a new deep learning approach for solving regularized unbalanced optimal transport (RUOT) and inferring continuous unbalanced stochastic dynamics from observed snapshots. Based on the RUOT form, our method models these dynamics without requiring prior knowledge of growth and death processes or additional information, allowing them to be learned directly from data. Theoretically, we explore the connections between the RUOT and Schrödinger bridge problem and discuss the key challenges and potential solutions. The effectiveness of our method is demonstrated with a synthetic gene regulatory network, high-dimensional Gaussian Mixture Model, and single-cell RNA-seq data from blood development. Compared with other methods, our approach accurately identifies growth and transition patterns, eliminates false transitions, and constructs the Waddington developmental landscape. Our code is available at: https://github.com/zhenyiizhang/DeepRUOT.

3037. Unlocking Point Processes through Point Set Diffusion

链接：https://iclr.cc/virtual/2025/poster/31003 abstract： Point processes model the distribution of random point sets in mathematical spaces, such as spatial and temporal domains, with applications in fields like seismology, neuroscience, and economics.Existing statistical and machine learning models for point processes are predominantly constrained by their reliance on the characteristic intensity function, introducing an inherent trade-off between efficiency and flexibility.In this paper, we introduce Point Set Diffusion, a diffusion-based latent variable model that can represent arbitrary point processes on general metric spaces without relying on the intensity function.By directly learning to stochastically interpolate between noise and data point sets, our approach effectively captures the distribution of point processes and enables efficient, parallel sampling and flexible generation for complex conditional tasks.Experiments on synthetic and real-world datasets demonstrate that Point Set Diffusion achieves state-of-the-art performance in unconditional and conditional generation of spatial and spatiotemporal point processes while providing up to orders of magnitude faster sampling.

3038. Learning Spatial-Semantic Features for Robust Video Object Segmentation

链接：https://iclr.cc/virtual/2025/poster/30411 abstract： Tracking and segmenting multiple similar objects with distinct or complex parts in long-term videos is particularly challenging due to the ambiguity in identifying target components and the confusion caused by occlusion, background clutter, and changes in appearance or environment over time. In this paper, we propose a robust video object segmentation framework that learns spatial-semantic features and discriminative object queries to address the above issues. Specifically, we construct a spatial-semantic block comprising a semantic embedding component and a spatial dependency modeling part for associating global semantic features and local spatial features, providing a comprehensive target representation. In addition, we develop a masked cross-attention module to generate object queries that focus on the most discriminative parts of target objects during query propagation, alleviating noise accumulation to ensure effective long-term query propagation. The experimental results show that the proposed method sets new state-of-the-art performance on multiple data sets, including the DAVIS2017 test (\textbf{87.8\%}), YoutubeVOS 2019 (\textbf{88.1\%}), MOSE val (\textbf{74.0\%}), and LVOS test (\textbf{73.0\%}), which demonstrate the effectiveness and generalization capacity of the proposed method. We will make all the source code and trained models publicly available.

3039. SafeDiffuser: Safe Planning with Diffusion Probabilistic Models

链接：https://iclr.cc/virtual/2025/poster/28682 abstract： Diffusion models have shown promise in data-driven planning. While these planners are commonly employed in applications where decisions are critical, they still lack established safety guarantees. In this paper, we address this limitation by introducing SafeDiffuser, a method to equip diffusion models with safety guarantees via control barrier functions. The key idea of our approach is to embed finite-time diffusion invariance, i.e., a form of specification consisting of safety constraints, into the denoising diffusion procedure. This way we enable data generation under safety constraints. We show that SafeDiffusers maintain the generative performance of diffusion models while also providing robustness in safe data generation. We evaluate our method on a series of tasks, including maze path generation, legged robot locomotion, and 3D space manipulation, and demonstrate the advantages of robustness over vanilla diffusion models.

3040. PseDet: Revisiting the Power of Pseudo Label in Incremental Object

Detection

链接：https://iclr.cc/virtual/2025/poster/30145 abstract： Incremental Objection Detection (IOD) facilitates the expansion of the usage scope of object detectors without forgetting previously acquired knowledge. Current approaches mostly adopt response-level knowledge distillation to overcome forgetting issues, by conducting implicit memory replay from the teacher model on new training data. However, this indirect learning paradigm does not fully leverage the knowledge generated by the teacher model. In this paper, we dive deeper into the mechanism of pseudo-labeling in incremental object detection by investigating three critical problems: (a) the upper bound quality of the pseudo labels is greatly limited by the previous model, (b) fixed score thresholds for label filtering, without considering the distribution across categories, and (c) the confidence score generated by the model does not well reflect the quality of the localization. Based on these observations, we propose a simple yet effective pseudo-labeling continual object detection framework, namely PseDet. Specifically, we introduce the spatio-temporal enhancement module to alleviate the negative effects when learning noisy data from the previous model. Considering the score distribution divergence across different classes, we propose the Categorical Adaptive Label Selector with a simple mathematical prior and fast K-Means pre-computation to dynamically determine the class-wise filtering threshold. In order to align the label score with the localization quality of the pseudo labels, we project the score through non-linear mapping to calibrate the distribution and integrate it into the new-step supervision. Extensive experiments on the competitive COCO benchmarks demonstrate the effectiveness and generalization of PseDet. Notably, it achieves 43.5+/41.2+ mAP under the 1/4-step incremental settings, achieving new state-of-the-art performance.

3041. The Breakdown of Gaussian Universality in Classification of High-dimensional Linear Factor Mixtures

链接：https://iclr.cc/virtual/2025/poster/29450 abstract： The assumption of Gaussian or Gaussian mixture data has been extensively exploited in a long series of precise performance analyses of machine learning (ML) methods, on large datasets having comparably numerous samples and features. To relax this restrictive assumption, subsequent efforts have been devoted to establish "Gaussian equivalent principles" by studying scenarios of Gaussian universality where the asymptotic performance of ML methods on non-Gaussian data remains unchanged when replaced with Gaussian data having the same mean and covariance.Beyond the realm of Gaussian universality, there are few exact results on how the data distribution affects the learning performance. In this article, we provide a precise high-dimensional characterization of empirical risk minimization, for classification under a general mixture data setting of linear factor models that extends Gaussian mixtures. The Gaussian universality is shown to break down under this setting, in the sense that the asymptotic learning performance depends on the data distribution beyond the class means and covariances.To clarify the limitations of Gaussian universality in the classification of mixture data and to understand the impact of its breakdown, we specify conditions for Gaussian universality and discuss their implications for the choice of loss function.

3042. Chain-of-Action: Faithful and Multimodal Question Answering through Large Language Models

链接：https://iclr.cc/virtual/2025/poster/31221 abstract： We present a Chain-of-Action (CoA) framework for multimodal and retrieval-augmented Question-Answering (QA). Compared to the literature, CoA overcomes two major challenges of current QA applications: (i) unfaithful hallucination that is inconsistent with real-time or domain facts and (ii) weak reasoning performance over compositional information. Our key contribution is a novel reasoning-retrieval mechanism that decomposes a complex question into a reasoning chain via systematic prompting and pre-designed actions. Methodologically, we propose three types of domain-adaptable `Plug-and-Play' actions for retrieving real-time information from heterogeneous sources. We also propose a multi-reference faith score to verify conflicts in the answers.In addition, our system demonstrates that detecting the knowledge boundaries of LLMs can significantly reduce both LLM interaction frequency and tokens usage in QA tasks. Empirically, we exploit both public benchmarks and a Web3 case study to demonstrate the capability of CoA over other methods.

3043. Web Agents with World Models: Learning and Leveraging Environment Dynamics in Web Navigation

链接：https://iclr.cc/virtual/2025/poster/28448 abstract： Large language models (LLMs) have recently gained much attention in building autonomous agents. However, performance of current LLM-based web agents in long-horizon tasks is far from optimal, often yielding errors such as repeatedly buying a non-refundable flight ticket. By contrast, humans can avoid such an irreversible mistake, as we have an awareness of the potential outcomes (e.g., losing money) of our actions, also known as the "world model". Motivated by this, our study first starts with preliminary analyses, confirming the absence of world models in current LLMs (e.g., GPT-4o, Claude-3.5-Sonnet, etc.). Then, we present a World-model-augmented (WMA) web agent, which simulates the outcomes of its actions for better decision-making. To overcome the challenges in training LLMs as world models predicting next observations, such as repeated elements across observations and long HTML inputs, we propose a transition-focused observation abstraction, where the prediction objectives are free-form natural language descriptions exclusively highlighting important state differences between time steps. Experiments on WebArena and Mind2Web show that our world models improve agents' policy selection without training and demonstrate our agents' cost- and time-efficiency compared to recent tree-search-based agents.

3044. NarrativeBridge: Enhancing Video Captioning with Causal-Temporal Narrative

链接：https://iclr.cc/virtual/2025/poster/29120 abstract： Existing video captioning benchmarks and models lack causal-temporal narrative, which is sequences of events linked through cause and effect, unfolding over time and driven by characters or agents. This lack of narrative restricts models' ability to generate text descriptions that capture the causal and temporal dynamics inherent in video content. To address this gap, we propose NarrativeBridge, an approach comprising of: (1) a novel Causal-Temporal Narrative (CTN) captions benchmark generated using a large language model and few-shot prompting, explicitly encoding cause-effect temporal relationships in video descriptions; and (2) a Cause-Effect Network (CEN) with separate encoders for capturing cause and effect dynamics, enabling effective learning and generation of captions with causal-temporal narrative. Extensive experiments demonstrate that CEN significantly outperforms state-of-the-art models in articulating the causal and temporal aspects of video content: 17.88 and 17.44 CIDEr on the MSVD-CTN and MSRVTT-CTN datasets, respectively. Cross-dataset evaluations further showcase CEN's strong generalization capabilities. The proposed framework understands and generates nuanced text descriptions with intricate causal-temporal narrative structures present in videos, addressing a critical limitation in video captioning. For project details, visit https://narrativebridge.github.io/.

3045. FedTMOS: Efficient One-Shot Federated Learning with Tsetlin Machine

链接：https://iclr.cc/virtual/2025/poster/31038 abstract： One-Shot Federated Learning (OFL) is a promising approach that reduce communication to a single round, minimizing latency and resource consumption. However, existing OFL methods often rely on Knowledge Distillation, which introduce server-side training, increasing latency. While neuron matching and model fusion techniques bypass server-side training, they struggle with alignment when heterogeneous data is present. To address these challenges, we proposed One-Shot Federated Learning with Tsetlin Machine (FedTMOS), a novel data-free OFL framework built upon the low-complexity and class-adaptive properties of the Tsetlin Machine. FedTMOS first clusters then reassigns class-specific weights to form models using an inter-class maximization approach, efficiently generating balanced server models without requiring additional training. Our extensive experiments demonstrate that FedTMOS significantly outperforms its ensemble counterpart by an average of 6.16%, and the leading state-of-the-art OFL baselines by 7.22% across various OFL settings. Moreover, FedTMOS achieves at least a $2.3\times$ reduction in upload communication costs and a $75\times$ reduction in server latency compared to methods requiring server-side training. These results establish FedTMOS as a highly efficient and practical solution for OFL scenarios.

3046. Asynchronous RLHF: Faster and More Efficient Off-Policy RL for Language Models

链接：https://iclr.cc/virtual/2025/poster/30332 abstract： The dominant paradigm for RLHF is *online* and *on-policy* RL: synchronously generating from the large language model (LLM) policy, labelling with a reward model, and learning using feedback on the LLM's own outputs. While performant, this paradigm is computationally inefficient. Inspired by classical deep RL literature, we propose separating generation and learning in RLHF. This enables asynchronous generation of new samples while simultaneously training on old samples, leading to faster training and more compute-optimal scaling. However, asynchronous training relies on an underexplored regime, online but *off-policy* RLHF: learning on samples from previous iterations of our model which give a worse training signal. We tackle the fundamental challenge in this regime: how much off-policyness can we tolerate for asynchronous training to speed up learning but maintain performance? Among several RLHF algorithms we test, online DPO is found to be most robust to off-policy data, and robustness increases with the scale of the policy model. We study further compute optimizations for asynchronous RLHF but find that they come at a performance cost, giving rise to a trade-off. We verify the scalability of asynchronous RLHF by training a general-purpose chatbot from LLaMA 3.1 8B on an instruction-following task $\sim$40\% faster than a synchronous run while matching final performance. Finally, we extend our results to math and reasoning to demonstrate asynchronous RL can finetune Rho 1B on GSM8k $\sim$70\% faster while matching synchronous accuracy.

3047. Everything is Editable: Extend Knowledge Editing to Unstructured Data in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29325 abstract： Recent knowledge editing methods have primarily focused on modifying structured knowledge in large language models. However, this task setting overlooks the fact that a significant portion of real-world knowledge is stored in an unstructured format, characterized by long-form content, noise, and a complex yet comprehensive nature.Techniques like "local layer key-value storage" and "term-driven optimization", as used in previous methods like MEMIT, are not effective for handling unstructured knowledge.To address these challenges, we propose a novel Unstructured Knowledge Editing method, namely UnKE, which extends previous assumptions in the layer dimension and token dimension.Firstly, in the layer dimension, we propose non-local block key-value storage to replace local layer key-value storage, increasing the representation ability of key-value pairs and incorporating attention layer knowledge. Secondly, in the token dimension, we replace "term-driven optimization" with "cause-driven optimization", which edits the last token directly while preserving context, avoiding the need to locate terms and preventing the loss of context information.Results on newly proposed unstructured knowledge editing dataset (UnKEBench) and traditional structured datasets demonstrate that UnKE achieves remarkable performance, surpassing strong baselines. In addition, UnKE has robust batch editing and sequential editing

capabilities.

3048. Learning View-invariant World Models for Visual Robotic Manipulation

链接：https://iclr.cc/virtual/2025/poster/27921 abstract： Robotic manipulation tasks often rely on visual inputs from cameras to perceive the environment. However, previous approaches still suffer from performance degradation when the camera’s viewpoint changes during manipulation. In this paper, we propose ReViWo (Representation learning for View-invariant World model), leveraging multi-view data to learn robust representations for control under viewpoint disturbance. ReViWo utilizes an autoencoder framework to reconstruct target images by an architecture that combines view-invariant representation (VIR) and view-dependent representation. To train ReViWo, we collect multi-view data in simulators with known view labels, meanwhile, ReViWo is simutaneously trained on Open X-Embodiment datasets without view labels. The VIR is then used to train a world model on pre-collected manipulation data and a policy through interaction with the world model. We evaluate the effectiveness of ReViWo in various viewpoint disturbance scenarios, including control under novel camera positions and frequent camera shaking, using the Meta-world & PandaGym environments. Besides, we also conduct experiments on real world ALOHA robot. The results demonstrate that ReViWo maintains robust performance under viewpoint disturbance, while baseline methods suffer from significant performance degradation. Furthermore, we show that the VIR captures task-relevant state information and remains stable for observations from novel viewpoints, validating the efficacy of the ReViWo approach.

3049. How Gradient descent balances features: A dynamical analysis for two-layer neural networks

链接：https://iclr.cc/virtual/2025/poster/31163 abstract： This paper investigates the fundamental regression task of learning k neurons (\emph{a.k.a.} teachers) from Gaussian input, using two-layer ReLU neural networks with width m (\emph{a.k.a.} students) and $m, k= \mathcal{O}(1)$, trained via gradient descent under proper initialization and a small step-size. Our analysis follows a three-phase structure: \emph{alignment} after weak recovery, \emph{tangential growth}, and \emph{local convergence}, providing deeper insights into the learning dynamics of gradient descent (GD). We prove the global convergence at the rate of $\mathcal{O}(T^{-3})$ for the zero loss of excess risk. Additionally, our results show that GD automatically groups and balances student neurons, revealing an implicit bias toward achieving the minimum ``balanced" ℓ_2-norm in the solution. Our work extends beyond previous studies in exact-parameterization setting ($m = k = 1$, (Yehudai and Ohad, 2020)) and single-neuron setting ($m \geq k = 1$, (Xu and Du, 2023)). The key technical challenge lies in handling the interactions between multiple teachers and students during training, which we address by refining the alignment analysis in Phase 1 and introducing a new dynamic system analysis for tangential components in Phase 2. Our results pave the way for further research on optimizing neural network training dynamics and understanding implicit biases in more complex architectures.

3050. A Unifying Framework for Representation Learning

链接：https://iclr.cc/virtual/2025/poster/29352 abstract： As the field of representation learning grows, there has been a proliferation of different loss functions to solve different classes of problems. We introduce a single information-theoretic equation that generalizes a large collection of mod- ern loss functions in machine learning. In particular, we introduce a framework that shows that several broad classes of machine learning methods are precisely minimizing an integrated KL divergence between two conditional distributions: the supervisory and learned representations. This viewpoint exposes a hidden information geometry underlying clustering, spectral methods, dimensionality re- duction, contrastive learning, and supervised learning. This framework enables the development of new loss functions by combining successful techniques from across the literature. We not only present a wide array of proofs, connecting over 23 different approaches, but we also leverage these theoretical results to create state-of-the-art unsupervised image classifiers that achieve a +8% improvement over the prior state-of-the-art on unsupervised classification on ImageNet-1K. We also demonstrate that I-Con can be used to derive principled debiasing methods which improve contrastive representation learners.

3051. Minimax Optimal Two-Stage Algorithm For Moment Estimation Under Covariate Shift

链接：https://iclr.cc/virtual/2025/poster/28348 abstract： Covariate shift occurs when the distribution of input features differs between the training and testing phases. In covariate shift, estimating an unknown function's moment is a classical problem that remains under-explored, despite its common occurrence in real-world scenarios. In this paper, we investigate the minimax lower bound of the problem when the source and target distributions are known. To achieve the minimax optimal bound (up to a logarithmic factor), we propose a two-stage algorithm. Specifically, it first trains an optimal estimator for the function under the source distribution, and then uses a likelihood ratio reweighting procedure to calibrate the moment estimator. In practice, the source and target distributions are typically unknown, and estimating the likelihood ratio may be unstable. To solve this problem, we propose a truncated version of the estimator that ensures double robustness and provide the corresponding upper bound. Extensive numerical studies on synthetic examples confirm our theoretical findings and further illustrate the effectiveness of our proposed method.

3052. TimeKAN: KAN-based Frequency Decomposition Learning Architecture for Long-term Time Series Forecasting

链接：https://iclr.cc/virtual/2025/poster/27844 abstract： Real-world time series often have multiple frequency components that are intertwined with each other, making accurate time series forecasting challenging. Decomposing the mixed frequency components into multiple single frequency components is a natural choice. However, the information density of patterns varies across different frequencies, and employing a uniform modeling approach for different frequency components can lead to inaccurate characterization. To address this challenges, inspired by the flexibility of the recent Kolmogorov-Arnold Network (KAN), we propose a KAN-based Frequency Decomposition Learning architecture (TimeKAN) to address the complex forecasting challenges caused by multiple frequency mixtures. Specifically, TimeKAN mainly consists of three components: Cascaded Frequency Decomposition (CFD) blocks, Multi-order KAN Representation Learning (M-KAN) blocks and Frequency Mixing blocks. CFD blocks adopt a bottom-up cascading approach to obtain series representations for each frequency band. Benefiting from the high flexibility of KAN, we design a novel M-KAN block to learn and represent specific temporal patterns within each frequency band. Finally, Frequency Mixing blocks is used to recombine the frequency bands into the original format. Extensive experimental results across multiple real-world time series datasets demonstrate that TimeKAN achieves state-of-the-art performance as an extremely lightweight architecture. Code is available at https://github.com/huangst21/TimeKAN.

3053. From Models to Microtheories: Distilling a Model's Topical Knowledge for Grounded Question-Answering

链接：https://iclr.cc/virtual/2025/poster/30107 abstract： Recent reasoning methods (e.g., chain-of-thought) help users understand how language models (LMs) answer a single question, but they do little to reveal the LM’s overall understanding, or “theory,” about the question’s topic, making it still hard to trust the model. Our goal is to materialize such theories - here called microtheories (a linguistic analog of logical microtheories) - as a set of sentences encapsulating an LM’s core knowledge about a topic. These statements systematically work together to entail answers to a set of questions to both engender trust and improve performance. Our approach is to first populate a knowledge store with (model-generated) sentences that entail answers to training questions, and then distill those down to a core microtheory which is concise, general, and non-redundant. We show that, when added to a general corpus (e.g., Wikipedia), microtheories can supply critical information not necessarily present in the corpus, improving both a model’s ability to ground its answers to verifiable knowledge (i.e., show how answers are systematically entailed by documents in the corpus, grounding up to +8% more answers), and the accuracy of those grounded answers (up to +8% absolute). We also show that, in a human evaluation in the medical domain, our distilled microtheories contain a significantly higher concentration of topically critical facts than the non-distilled knowledge store. Finally, we show we can quantify the coverage of a microtheory for a topic (characterized by a dataset) using a notion of p-relevance. Together, these suggest that microtheories are an efficient distillation of an LM’s topic-relevant knowledge, that they can usefully augment existing corpora, and can provide both performance gains and an interpretable, verifiable window into the model’s knowledge of a topic.

3054. Exploring Local Memorization in Diffusion Models via Bright Ending Attention

链接：https://iclr.cc/virtual/2025/poster/28321 abstract： Text-to-image diffusion models have achieved unprecedented proficiency in generating realistic images. However, their inherent tendency to memorize and replicate training data during inference raises significant concerns, including potential copyright infringement. In response, various methods have been proposed to evaluate, detect, and mitigate memorization. Our analysis reveals that existing approaches significantly underperform in handling local memorization, where only specific image regions are memorized, compared to global memorization, where the entire image is replicated. Also, they cannot locate the local memorization regions, making it hard to investigate locally. To address these, we identify a novel "bright ending" (BE) anomaly in diffusion models prone to memorizing training images. BE refers to a distinct cross-attention pattern observed in text-to-image diffusion models, where memorized image patches exhibit significantly greater attention to the final text token during the last inference step than non-memorized patches. This pattern highlights regions where the generated image replicates training data and enables efficient localization of memorized regions. Equipped with this, we propose a simple yet effective method to integrate BE into existing frameworks, significantly improving their performance by narrowing the performance gap caused by local memorization. Our results not only validate the successful execution of the new localization task but also establish new state-of-the-art performance across all existing tasks, underscoring the significance of the BE phenomenon.

3055. ALBAR: Adversarial Learning approach to mitigate Biases in Action Recognition

链接：https://iclr.cc/virtual/2025/poster/30700 abstract： Bias in machine learning models can lead to unfair decision making, and while it has been well-studied in the image and text domains, it remains underexplored in action recognition. Action recognition models often suffer from background bias (i.e., inferring actions based on background cues) and foreground bias (i.e., relying on subject appearance), which can be detrimental to real-life applications such as autonomous vehicles or assisted living monitoring. While prior approaches have mainly focused on mitigating background bias using specialized augmentations, we thoroughly study both foreground and background bias. We propose ALBAR, a novel adversarial training method that mitigates foreground and background biases without requiring specialized knowledge of the bias attributes. Our framework applies an adversarial cross-entropy loss to the sampled static clip (where all the frames are the same) and aims to make its class probabilities uniform using a proposed entropy maximization loss. Additionally, we introduce a gradient penalty loss for regularization against the debiasing process. We evaluate our method on established background and foreground bias

protocols, setting a new state-of-the-art and strongly improving combined debiasing performance by over 12% absolute on HMDB51. Furthermore, we identify an issue of background leakage in the existing UCF101 protocol for bias evaluation which provides a shortcut to predict actions and does not provide an accurate measure of the debiasing capability of a model. We address this issue by proposing more fine-grained segmentation boundaries for the actor, where our method also outperforms existing approaches.

3056. Analytic DAG Constraints for Differentiable DAG Learning

链接：https://iclr.cc/virtual/2025/poster/28370 abstract： Recovering the underlying Directed Acyclic Graph (DAG) structures from observational data presents a formidable challenge, partly due to the combinatorial nature of the DAG-constrained optimization problem. Recently, researchers have identified gradient vanishing as one of the primary obstacles in differentiable DAG learning and have proposed several DAG constraints to mitigate this issue. By developing the necessary theory to establish a connection between analytic functions and DAG constraints, we demonstrate that analytic functions from the set $\{f(x) = c_0 + \sum_{i=1}^{\infty}c_ix^i | \forall i > 0, c_i > 0; r = \lim_{i\rightarrow \infty}c_{i}/c_{i+1} > 0\}$ can be employed to formulate effective DAG constraints. Furthermore, we establish that this set of functions is closed under several functional operators, including differentiation, summation, and multiplication. Consequently, these operators can be leveraged to create novel DAG constraints based on existing ones. Using these properties, we design a series of DAG constraints and develop an efficient algorithm to evaluate them. Experiments in various settings demonstrate that our DAG constraints outperform previous state-of-the-art comparators. Our implementation is available at https://github.com/zzhang1987/AnalyticDAGLearning.

3057. Artificial Kuramoto Oscillatory Neurons

链接：https://iclr.cc/virtual/2025/poster/28387 abstract： It has long been known in both neuroscience and AI that ``binding'' between neurons leads to a form of competitive learning where representations are compressed in order to represent more abstract concepts in deeper layers of the network. More recently, it was also hypothesized that dynamic (spatiotemporal) representations play an important role in both neuroscience and AI. Building on these ideas, we introduce Artificial Kuramoto Oscillatory Neurons (AKOrN) as a dynamical alternative to threshold units, which can be combined with arbitrary connectivity designs such as fully connected, convolutional, or attentive mechanisms. Our generalized Kuramoto updates bind neurons together through their synchronization dynamics. We show that this idea provides performance improvements across a wide spectrum of tasks such as unsupervised object discovery, adversarial robustness, calibrated uncertainty quantification, and reasoning. We believe that these empirical results show the importance of rethinking our assumptions at the most basic neuronal level of neural representation, and in particular show the importance of dynamical representations.

3058. InstructRAG: Instructing Retrieval-Augmented Generation via Self-Synthesized Rationales

链接：https://iclr.cc/virtual/2025/poster/29783 abstract： Retrieval-augmented generation (RAG) has shown promising potential to enhance the accuracy and factuality of language models (LMs). However, imperfect retrievers or noisy corpora can introduce misleading or even erroneous information to the retrieved contents, posing a significant challenge to the generation quality. Existing RAG methods typically address this challenge by directly predicting final answers despite potentially noisy inputs, resulting in an implicit denoising process that is difficult to interpret and verify. On the other hand, the acquisition of explicit denoising supervision is often costly, involving significant human efforts. In this work, we propose InstructRAG, where LMs explicitly learn the denoising process through self-synthesized rationales --- First, we instruct the LM to explain how the ground-truth answer is derived from retrieved documents. Then, these rationales can be used either as demonstrations for in-context learning of explicit denoising or as supervised fine-tuning data to train the model. Compared to standard RAG approaches, InstructRAG requires no additional supervision, allows for easier verification of the predicted answers, and effectively improves generation accuracy. Experiments show InstructRAG consistently outperforms existing RAG methods in both training-free and trainable scenarios, achieving a relative improvement of 8.3% over the best baseline method on average across five knowledge-intensive benchmarks. Extensive analysis indicates that InstructRAG scales well with increased numbers of retrieved documents and consistently exhibits robust denoising ability even in out-of-domain datasets, demonstrating strong generalizability.

3059. Exploiting Structure in Offline Multi-Agent RL: The Benefits of Low Interaction Rank

链接：https://iclr.cc/virtual/2025/poster/30636 abstract： We study the problem of learning an approximate equilibrium in the offline multi-agent reinforcement learning (MARL) setting. We introduce a structural assumption---the interaction rank---and establish that functions with low interaction rank are significantly more robust to distribution shift compared to general ones. Leveraging this observation, we demonstrate that utilizing function classes with low interaction rank, when combined with regularization and no-regret learning, admits decentralized, computationally and statistically efficient learning in cooperative and competitive offline MARL. Our theoretical results are complemented by experiments that showcase the potential of critic architectures with low interaction rank in offline MARL, contrasting with commonly used single-agent value decomposition architectures.

3060. Towards Generalization Bounds of GCNs for Adversarially Robust Node Classification

链接：https://iclr.cc/virtual/2025/poster/29027 abstract： Adversarially robust generalization of Graph Convolutional Networks (GCNs) has garnered significant attention in various security-sensitive application areas, driven by intrinsic adversarial vulnerability. Albeit remarkable empirical advancement, theoretical understanding of the generalization behavior of GCNs subjected to adversarial attacks remains elusive. To make progress on the mystery, we establish unified high-probability generalization bounds for GCNs in the context of node classification, by leveraging adversarial Transductive Rademacher Complexity (TRC) and developing a novel contraction technique on graph convolution. Our bounds capture the interaction between generalization error and adversarial perturbations, revealing the importance of key quantities in mitigating the negative effects of perturbations, such as low-dimensional feature projection, perturbation-dependent norm regularization, normalized graph matrix, proper number of network layers, etc. Furthermore, we provide TRC-based bounds of popular GCNs with ℓ_r-norm-additive perturbations for arbitrary $r\geq 1$. A comparison of theoretical results demonstrates that specific network architectures (e.g., residual connection) can help alleviate the cumulative effect of perturbations during the forward propagation of deep GCNs. Experimental results on benchmark datasets validate our theoretical findings.

3061. Physics-informed Temporal Difference Metric Learning for Robot Motion Planning

链接：https://iclr.cc/virtual/2025/poster/29546 abstract： The motion planning problem involves finding a collision-free path from a robot's starting to its target configuration. Recently, self-supervised learning methods have emerged to tackle motion planning problems without requiring expensive expert demonstrations. They solve the Eikonal equation for training neural networks and lead to efficient solutions. However, these methods struggle in complex environments because they fail to maintain key properties of the Eikonal equation, such as optimal value functions and geodesic distances. To overcome these limitations, we propose a novel self-supervised temporal difference metric learning approach that solves the Eikonal equation more accurately and enhances performance in solving complex and unseen planning tasks. Our method enforces Bellman's principle of optimality over finite regions, using temporal difference learning to avoid spurious local minima while incorporating metric learning to preserve the Eikonal equation's essential geodesic properties. We demonstrate that our approach significantly outperforms existing self-supervised learning methods in handling complex environments and generalizing to unseen environments, with robot configurations ranging from 2 to 12 degrees of freedom (DOF).

3062. Scalable and Certifiable Graph Unlearning: Overcoming the Approximation Error Barrier

链接：https://iclr.cc/virtual/2025/poster/28301 abstract： Graph unlearning has emerged as a pivotal research area for ensuring privacy protection, given the widespread adoption of Graph Neural Networks (GNNs) in applications involving sensitive user data. Among existing studies, certified graph unlearning is distinguished by providing robust privacy guarantees. However, current certified graph unlearning methods are impractical for large-scale graphs because they necessitate the costly re-computation of graph propagation for each unlearning request. Although numerous scalable techniques have been developed to accelerate graph propagation for GNNs, their integration into certified graph unlearning remains uncertain as these scalable approaches introduce approximation errors into node embeddings. In contrast, certified graph unlearning demands bounded model error on exact node embeddings to maintain its certified guarantee. To address this challenge, we present ScaleGUN, the first approach to scale certified graph unlearning to billion-edge graphs. ScaleGUN integrates the approximate graph propagation technique into certified graph unlearning, offering certified guarantees for three unlearning scenarios: node feature, edge and node unlearning. Extensive experiments on real-world datasets demonstrate the efficiency and unlearning efficacy of ScaleGUN. Remarkably, ScaleGUN accomplishes $(\epsilon,\delta)=(1,10^{-4})$ certified unlearning on the billion-edge graph ogbn-papers100M in 20 seconds for a 5,000 random edge removal request -- of which only 5 seconds are required for updating the node embeddings -- compared to 1.91 hours for retraining and 1.89 hours for re-propagation. Our code is available at https://github.com/luyi256/ScaleGUN.

3063. Restating the Proof of Linear Convergence for Linear GNNs

链接：https://iclr.cc/virtual/2025/poster/31357 abstract： We lead the readers through the core proof of a pioneering paper that studies the training dynamics of linear GNNs. First, we reorganize the proof and provide a more concise and reader-friendly version, highlighting several key components. In doing so, we identify a hidden error and correct it, demonstrating that it has no impact on the main result. Additionally, we offer a dialectical discussion on the strengths and an overlooked aspect of the approach.

3064. TGB-Seq Benchmark: Challenging Temporal GNNs with Complex Sequential Dynamics

链接：https://iclr.cc/virtual/2025/poster/30756 abstract： Future link prediction is a fundamental challenge in various real-world dynamic systems. To address this, numerous temporal graph neural networks (temporal GNNs) and benchmark datasets have

been developed. However, these datasets often feature excessive repeated edges and lack complex sequential dynamics, a key characteristic inherent in many real-world applications such as recommender systems and "Who-To-Follow" on social networks. This oversight has led existing methods to inadvertently downplay the importance of learning sequential dynamics, focusing primarily on predicting repeated edges.In this study, we demonstrate that existing methods, such as GraphMixer and DyGFormer, are inherently incapable of learning simple sequential dynamics, such as "a user who has followed OpenAI and Anthropic is more likely to follow AI at Meta next." Motivated by this issue, we introduce the Temporal Graph Benchmark with Sequential Dynamics (TGB-Seq), a new benchmark carefully curated to minimize repeated edges, challenging models to learn sequential dynamics and generalize to unseen edges. TGB-Seq comprises large real-world datasets spanning diverse domains, including e-commerce interactions, movie ratings, business reviews, social networks, citation networks and web link networks. Benchmarking experiments reveal that current methods usually suffer significant performance degradation and incur substantial training costs on TGB-Seq, posing new challenges and opportunities for future research. TGB-Seq datasets, leaderboards, and example codes are available at https://tgb-seq.github.io/.

3065. Energy-based Backdoor Defense Against Federated Graph Learning

链接：https://iclr.cc/virtual/2025/poster/30953 abstract： Federated Graph Learning is rapidly evolving as a privacy-preserving collaborative approach. However, backdoor attacks are increasingly undermining federated systems by injecting carefully designed triggers that lead to the model making incorrect predictions. Trigger structures and injection locations in Federated Graph Learning are more diverse, making traditional federated defense methods less effective. In our work, we propose an effective Federated Graph Backdoor Defense using Topological Graph Energy (FedTGE). At the local client level, it injects distribution knowledge into the local model, assigning low energy to benign samples and high energy to the constructed malicious substitutes, and selects benign clients through clustering. At the global server level, the energy elements uploaded by each client are treated as new nodes to construct a global energy graph for energy propagation, making the selected clients' energy elements more similar and further adjusting the aggregation weights. Our method can handle high data heterogeneity, does not require a validation dataset, and is effective under both small and large malicious proportions. Extensive results on various settings of federated graph scenarios under backdoor attacks validate the effectiveness of this approach.

3066. How Two-Layer Neural Networks Learn, One (Giant) Step at a Time

链接：https://iclr.cc/virtual/2025/poster/31384 abstract： For high-dimensional Gaussian data, we investigate theoretically how the features of a two-layer neural network adapt to the structure of the target function through a few large batch gradient descent steps, leading to an improvement in the approximation capacity with respect to the initialization. First, we compare the influence of batch size to that of multiple (but finitely many) steps. For a single gradient step, a batch of size $n = O(d)$ is both necessary and sufficient to align with the target function, although only a single direction can be learned. In contrast, $n = O(d^2)$ is essential for neurons to specialize in multiple relevant directions of the target with a single gradient step. Even in this case, we show there might exist ``hard'' directions requiring $n = O(d^\ell)$ samples to be learned, where ℓ is known as the leap index of the target. Second, we show that the picture drastically improves over multiple gradient steps: a batch size of $n = O(d)$ is indeed sufficient to learn multiple target directions satisfying a staircase property, where more and more directions can be learned over time. Finally, we discuss how these directions allow for a drastic improvement in the approximation capacity and generalization error over the initialization, illustrating a separation of scale between the random features/lazy regime and the feature learning regime. Our technical analysis leverages a combination of techniques related to concentration, projection-based conditioning, and Gaussian equivalence, which we believe are of independent interest. By pinning down the conditions necessary for specialization and learning, our results highlight the intertwined role of the structure of the task to learn, the detail of the algorithm (the batch size), and the architecture (i.e., the number of hidden neurons), shedding new light on how neural networks adapt to the feature and learn complex task from data over time.

3067. Mechanistic Permutability: Match Features Across Layers

链接：https://iclr.cc/virtual/2025/poster/29946 abstract： Understanding how features evolve across layers in deep neural networks is a fundamental challenge in mechanistic interpretability, particularly due to polysemanticity and feature superposition. While Sparse Autoencoders (SAEs) have been used to extract interpretable features from individual layers, aligning these features across layers has remained an open problem. In this paper, we introduce SAE Match, a novel, data-free method for aligning SAE features across different layers of a neural network. Our approach involves matching features by minimizing the mean squared error between the folded parameters of SAEs, a technique that incorporates activation thresholds into the encoder and decoder weights to account for differences in feature scales. Through extensive experiments on the Gemma 2 language model, we demonstrate that our method effectively captures feature evolution across layers, improving feature matching quality. We also show that features persist over several layers and that our approach can approximate hidden states across layers. Our work advances the understanding of feature dynamics in neural networks and provides a new tool for mechanistic interpretability studies.

3068. Moral Alignment for LLM Agents

链接：https://iclr.cc/virtual/2025/poster/29923 abstract： Decision-making agents based on pre-trained Large Language Models (LLMs) are increasingly being deployed across various domains of human activity. While their applications are currently rather specialized, several research efforts are underway to develop more generalist agents. As LLM-based systems become more agentic, their influence on human activity will grow and their transparency will decrease. Consequently, developing effective

methods for aligning them to human values is vital. The prevailing practice in alignment often relies on human preference data (e.g., in RLHF or DPO), in which values are implicit, opaque and are essentially deduced from relative preferences over different model outputs. In this work, instead of relying on human feedback, we introduce the design of reward functions that explicitly and transparently encode core human values for Reinforcement Learning-based fine-tuning of foundation agent models. Specifically, we use intrinsic rewards for the moral alignment of LLM agents. We evaluate our approach using the traditional philosophical frameworks of Deontological Ethics and Utilitarianism, quantifying moral rewards for agents in terms of actions and consequences on the Iterated Prisoner's Dilemma (IPD) environment. We also show how moral fine-tuning can be deployed to enable an agent to unlearn a previously developed selfish strategy. Finally, we find that certain moral strategies learned on the IPD game generalize to several other matrix game environments. In summary, we demonstrate that fine-tuning with intrinsic rewards is a promising general solution for aligning LLM agents to human values, and it might represent a more transparent and cost-effective alternative to currently predominant alignment techniques.

3069. Ask, and it shall be given: On the Turing completeness of prompting

链接：https://iclr.cc/virtual/2025/poster/30632 abstract： Since the success of GPT, large language models (LLMs) have revolutionized machine learning and have initiated the so-called LLM prompting paradigm. In the era of LLMs, people train a single general-purpose LLM and provide the LLM with different prompts to perform different tasks. However, such empirical success largely lacks theoretical understanding. Here, we present the first theoretical study on the LLM prompting paradigm to the best of our knowledge. In this work, we show that prompting is in fact Turing-complete: there exists a finite-size Transformer such that for any computable function, there exists a corresponding prompt following which the Transformer computes the function. Furthermore, we show that even though we use only a single finite-size Transformer, it can still achieve nearly the same complexity bounds as that of the class of all unbounded-size Transformers. Overall, our result reveals that prompting can enable a single finite-size Transformer to be efficiently universal, which establishes a theoretical underpinning for prompt engineering in practice.

3070. One-for-All Few-Shot Anomaly Detection via Instance-Induced Prompt Learning

链接：https://iclr.cc/virtual/2025/poster/29190 abstract： Anomaly detection methods under the 'one-for-all' paradigm aim to develop a unified model capable of detecting anomalies across multiple classes. However, these approaches typically require a large number of normal samples for model training, which may not always be feasible in practice. Few-shot anomaly detection methods can address scenarios with limited data but often require a tailored model for each class, struggling within the 'one-for-one' paradigm. In this paper, we first proposed the one-for-all few-shot anomaly detection method with the assistance of vision-language model. Different from previous CLIP-based methods learning fix prompts for each class, our method learn a class-shared prompt generator to adaptively generate suitable prompt for each instance. The prompt generator is trained by aligning the prompts with the visual space and utilizing guidance from general textual descriptions of normality and abnormality. Furthermore, we address the mismatch problem of the memory bank within one-for-all paradigm. Extensive experimental results on MVTec and VisA demonstrate the superiority of our method in few-shot anomaly detection task under the one-for-all paradigm.

3071. From Artificial Needles to Real Haystacks: Improving Retrieval Capabilities in LLMs by Finetuning on Synthetic Data

链接：https://iclr.cc/virtual/2025/poster/30747 abstract： Recent studies have shown that Large Language Models (LLMs) struggle to accurately retrieve information and maintain reasoning capabilities when processing long-context inputs. To address these limitations, we propose a finetuning approach utilizing a carefully designed synthetic dataset comprising numerical key-value retrieval tasks. Our experiments on models like GPT-3.5 Turbo and Mistral 7B demonstrate that finetuning LLMs on this dataset significantly improves LLMs' information retrieval and reasoning capabilities in longer-context settings. We present an analysis of the finetuned models, illustrating the transfer of skills from synthetic to real task evaluations (e.g., 10.5% improvement on 20 documents MDQA at position 10 for GPT-3.5 Turbo). We also find that finetuned LLMs' performance on general benchmarks remains almost constant while LLMs finetuned on other baseline long-context augmentation data can encourage hallucination (e.g., on TriviaQA, Mistral 7B finetuned on our synthetic data cause no performance drop while other baseline data can cause a drop that ranges from 2.33% to 6.19%). Our study highlights the potential of finetuning on synthetic data for improving the performance of LLMs on longer-context tasks.

3072. Process Reward Model with Q-value Rankings

链接：https://iclr.cc/virtual/2025/poster/27847 abstract： Process Reward Modeling (PRM) is critical for complex reasoning and decision-making tasks where the accuracy of intermediate steps significantly influences the overall outcome. Existing PRM approaches, primarily framed as classification problems, employ cross-entropy loss to independently evaluate each step's correctness. This method can lead to suboptimal reward distribution and does not adequately address the interdependencies among steps. To address these limitations, we introduce the Process Q-value Model (PQM), a novel framework that redefines PRM in the context of a Markov Decision Process. PQM optimizes Q-value rankings based on a novel comparative loss function, enhancing the model's ability to capture the intricate dynamics among sequential decisions. This approach provides a more granular and theoretically grounded methodology for process rewards. Our extensive empirical evaluations across various

sampling policies, language model backbones, and multi-step reasoning benchmarks show that PQM outperforms classification-based PRMs. The effectiveness of the comparative loss function is highlighted in our comprehensive ablation studies, confirming PQM's practical efficacy and theoretical advantage.

3073. OmniEdit: Building Image Editing Generalist Models Through Specialist Supervision

链接：https://iclr.cc/virtual/2025/poster/30213 abstract： Instruction-guided image editing methods have demonstrated significant potential by training diffusion models on automatically synthesized or manually annotated image editing pairs. However, these methods remain far from practical, real-life applications. We identify three primary challenges contributing to this gap. Firstly, existing models have limited editing skills due to the biased synthesis process. Secondly, these methods are trained with datasets with a high volume of noise and artifacts. This is due to the application of simple filtering methods like CLIP-score. Thirdly, all these datasets are restricted to a single low resolution and fixed aspect ratio, limiting the versatility to handle real-world use cases.In this paper, we present OmniEdit, which is an omnipotent editor to handle seven different image editing tasks with any aspect ratio seamlessly. Our contribution is in four folds: (1) OmniEdit is trained by utilizing the supervision from seven different specialist models to ensure task coverage. (2) we utilize importance sampling based on the scores provided by large multimodal models (like GPT-4o) instead of CLIP-score to improve the data quality. (3) we propose a new editing architecture called EditNet to greatly boost the editing success rate, (4) we provide images with different aspect ratios to ensure that our model can handle any image in the wild. We have curated a test set containing images of different aspect ratios, accompanied by diverse instructions to cover different tasks. Both automatic evaluation and human evaluations demonstrate that OmniEdit can significantly outperforms all the existing models.

3074. UniCon: Unidirectional Information Flow for Effective Control of Large-Scale Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/27986 abstract： We introduce UniCon, a novel architecture designed to enhance control and efficiency in training adapters for large-scale diffusion models. Unlike existing methods that rely on bidirectional interaction between the diffusion model and control adapter, UniCon implements a unidirectional flow from the diffusion network to the adapter, allowing the adapter alone to generate the final output. UniCon reduces computational demands by eliminating the need for the diffusion model to compute and store gradients during adapter training. Our results indicate that UniCon reduces GPU memory usage by one-third and increases training speed by 2.3 times, while maintaining the same adapter parameter size. Additionally, without requiring extra computational resources, UniCon enables the training of adapters with double the parameter volume of existing ControlNets. In a series of image conditional generation tasks, UniCon has demonstrated precise responsiveness to control inputs and exceptional generation capabilities.

3075. Efficient Cross-Episode Meta-RL

链接：https://iclr.cc/virtual/2025/poster/29485 abstract： We introduce Efficient Cross-Episodic Transformers (ECET), a new algorithm for online Meta-Reinforcement Learning that addresses the challenge of enabling reinforcement learning agents to perform effectively in previously unseen tasks. We demonstrate how past episodes serve as a rich source of in-context information, which our model effectively distills and applies to new contexts. Our learned algorithm is capable of outperforming the previous state-of-the-art and provides more efficient meta-training while significantly improving generalization capabilities. Experimental results, obtained across various simulated tasks of the MuJoCo, Meta-World and ManiSkill benchmarks, indicate a significant improvement in learning efficiency and adaptability compared to the state-of-the-art. Our approach enhances the agent's ability to generalize from limited data and paves the way for more robust and versatile AI systems.

3076. Do Vision-Language Models Represent Space and How? Evaluating Spatial Frame of Reference under Ambiguities

链接：https://iclr.cc/virtual/2025/poster/30786 abstract： Spatial expressions in situated communication can be ambiguous, as their meanings vary depending on the frames of reference (FoR) adopted by speakers and listeners. While spatial language understanding and reasoning by vision-language models (VLMs) have gained increasing attention, potential ambiguities in these models are still under-explored. To address this issue, we present the COnsistent Multilingual Frame Of Reference Test (COMFORT), an evaluation protocol to systematically assess the spatial reasoning capabilities of VLMs. We evaluate nine state-of-the-art VLMs using COMFORT. Despite showing some alignment with English conventions in resolving ambiguities, our experiments reveal significant shortcomings of VLMs: notably, the models (1) exhibit poor robustness and consistency, (2) lack the flexibility to accommodate multiple FoRs, and (3) fail to adhere to language-specific or culture-specific conventions in cross-lingual tests, as English tends to dominate other languages. With a growing effort to align vision-language models with human cognitive intuitions, we call for more attention to the ambiguous nature and cross-cultural diversity of spatial reasoning.

3077. COMBO: Compositional World Models for Embodied Multi-Agent Cooperation

链接：https://iclr.cc/virtual/2025/poster/29260 abstract： In this paper, we investigate the problem of embodied multi-agent cooperation, where decentralized agents must cooperate given only egocentric views of the world. To effectively plan in this setting, in contrast to learning world dynamics in a single-agent scenario, we must simulate world dynamics conditioned on an arbitrary number of agents' actions given only partial egocentric visual observations of the world. To address this issue of partial observability, we first train generative models to estimate the overall world state given partial egocentric observations. To enable accurate simulation of multiple sets of actions on this world state, we then propose to learn a compositional world model for multi-agent cooperation by factorizing the naturally composable joint actions of multiple agents and compositionally generating the video conditioned on the world state. By leveraging this compositional world model, in combination with Vision Language Models to infer the actions of other agents, we can use a tree search procedure to integrate these modules and facilitate online cooperative planning. We evaluate our methods on three challenging benchmarks with 2-4 agents. The results show our compositional world model is effective and the framework enables the embodied agents to cooperate efficiently with different agents across various tasks and an arbitrary number of agents, showing the promising future of our proposed methods. More videos can be found at \url{https://embodied-agi.cs.umass.edu/combo/}.

3078. Dynamic-LLaVA: Efficient Multimodal Large Language Models via Dynamic Vision-language Context Sparsification

链接：https://iclr.cc/virtual/2025/poster/28727 abstract： Multimodal Large Language Models (MLLMs) have achieved remarkable success in vision understanding, reasoning, and interaction. However, the inference computation and memory increase progressively with the generation of output tokens during decoding, directly affecting the efficacy of MLLMs. Existing methods attempt to reduce the vision context redundancy to achieve efficient MLLMs. Unfortunately, the efficiency benefits of the vision context reduction in the prefill stage gradually diminish during the decoding stage. To address this problem, we proposed a dynamic vision-language context sparsification framework Dynamic-LLaVA, which dynamically reduces the redundancy of vision context in the prefill stage and decreases the memory and computation overhead of the generated language context during decoding. Dynamic-LLaVA designs a tailored sparsification inference scheme for different inference modes, i.e., prefill, decoding with and without KV cache, to achieve efficient inference of MLLMs. In practice, Dynamic-LLaVA can reduce computation consumption by $\sim$75\% in the prefill stage. Meanwhile, throughout the entire generation process of MLLMs, Dynamic-LLaVA reduces the $\sim$50\% computation consumption under decoding without KV cache, while saving $\sim$50\% GPU memory overhead when decoding with KV cache, due to the vision-language context sparsification. Extensive experiments also demonstrate that Dynamic-LLaVA achieves efficient inference for MLLMs with negligible understanding and generation ability degradation or even performance gains compared to the full-context inference baselines. Code is available at https://github.com/Osilly/dynamic_llava.

3079. Learn Your Reference Model for Real Good Alignment

链接：https://iclr.cc/virtual/2025/poster/30247 abstract： Despite the fact that offline methods for Large Language Models (LLMs) alignment do not require a direct reward model, they remain susceptible to overoptimization. This issue arises when the trained model deviates excessively from the reference policy, leading to a decrease in sample quality. We propose a novel approach of offline alignment methods, called Trust Region (including variants TR-DPO, TR-IPO, TR-KTO), which dynamically updates the reference policy throughout the training process. Our results show that TR alignment methods effectively mitigate overoptimization, enabling models to maintain strong performance even when substantially deviating from the initial reference policy. We demonstrate the efficacy of these approaches not only through toy examples that exhibit reduced overoptimization, but also through direct, side-by-side comparisons in specific tasks such as helpful and harmless dialogue, as well as summarization, where they surpass conventional methods. Additionally, we report significant improvements in general-purpose assistant setups with the Llama3 model on the AlpacaEval 2 and Arena-Hard benchmarks, highlighting the advantages of Trust Region methods over classical approaches.

3080. Steering Protein Family Design through Profile Bayesian Flow

链接：https://iclr.cc/virtual/2025/poster/29753 abstract： Protein family design emerges as a promising alternative by combining the advantages of de novo protein design and mutation-based directed evolution.In this paper, we propose ProfileBFN, the Profile Bayesian Flow Networks, for specifically generative modeling of protein families. ProfileBFN extends the discrete Bayesian Flow Network from an MSA profile perspective, which can be trained on single protein sequences by regarding it as a degenerate profile, thereby achieving efficient protein family design by avoiding large-scale MSA data construction and training. Empirical results show that ProfileBFN has a profound understanding of proteins. When generating diverse and novel family proteins, it can accurately capture the structural characteristics of the family. The enzyme produced by this method is more likely than the previous approach to have the corresponding function, offering better odds of generating diverse proteins with the desired functionality.

3081. C-CLIP: Multimodal Continual Learning for Vision-Language Model

链接：https://iclr.cc/virtual/2025/poster/28116 abstract： Multimodal pre-trained models like CLIP need large image-text pairs for training but often struggle with domain-specific tasks. Since retraining with specialized and historical data incurs significant memory and time costs, it is important to continually learn new domains in the open world while preserving original performance. However, current continual learning research mainly focuses on single-modal scenarios, and the evaluation criteria are insufficient without considering image-text matching performance and the forgetting of zero-shot performance. This work

introduces image-caption datasets from various domains and establishes a multimodal vision-language continual learning benchmark. Then, a novel framework named C-CLIP is proposed, which not only prevents forgetting but also enhances new task learning impressively. Comprehensive experiments demonstrate that our method has strong continual learning ability across different domain image-text datasets, and has little forgetting of the original capabilities of zero-shot prediction, significantly outperforming existing methods.

3082. Local-Prompt: Extensible Local Prompts for Few-Shot Out-of-Distribution Detection

链接：https://iclr.cc/virtual/2025/poster/30380 abstract： Out-of-Distribution (OOD) detection, aiming to distinguish outliers from known categories, has gained prominence in practical scenarios. Recently, the advent of vision-language models (VLM) has heightened interest in enhancing OOD detection for VLM through few-shot tuning. However, existing methods mainly focus on optimizing global prompts, ignoring refined utilization of local information with regard to outliers. Motivated by this, we freeze global prompts and introduce Local-Prompt, a novel coarse-to-fine tuning paradigm to emphasize regional enhancement with local prompts. Our method comprises two integral components: global prompt guided negative augmentation and local prompt enhanced regional regularization. The former utilizes frozen, coarse global prompts as guiding cues to incorporate negative augmentation, thereby leveraging local outlier knowledge. The latter employs trainable local prompts and a regional regularization to capture local information effectively, aiding in outlier identification. We also propose regional-related metric to empower the enrichment of OOD detection. Moreover, since our approach explores enhancing local prompts only, it can be seamlessly integrated with trained global prompts during inference to boost the performance. Comprehensive experiments demonstrate the effectiveness and potential of our method. Notably, our method reduces average FPR95 by 5.17% against state-of-the-art method in 4-shot tuning on challenging ImageNet-1k dataset, even outperforming 16-shot results of previous methods.

3083. Adaptive Batch Size for Privately Finding Second-Order Stationary Points

链接：https://iclr.cc/virtual/2025/poster/28679 abstract： There is a gap between finding a first-order stationary point (FOSP) and a second-order stationary point (SOSP) under differential privacy constraints, and it remains unclear whether privately finding an SOSP is more challenging than finding an FOSP. Specifically, Ganesh et al. (2023) claimed that an α-SOSP can be found with $\alpha=\Tilde{O}(\frac{1}{n^{1/3}}+(\frac{\sqrt{d}}{n\epsilon})^{3/7})$, where n is the dataset size, d is the dimension, and ϵ is the differential privacy parameter.However, a recent analysis revealed an issue in their saddle point escape procedure, leading to weaker guarantees. Building on the SpiderBoost algorithm framework, we propose a new approach that uses adaptive batch sizes and incorporates the binary tree mechanism.Our method not only corrects this issue but also improves the results for privately finding an SOSP, achieving $\alpha=\Tilde{O}(\frac{1}{n^{1/3}}+(\frac{\sqrt{d}}{n\epsilon})^{1/2})$. This improved bound matches the state-of-the-art for finding a FOSP, suggesting that privately finding an SOSP may be achievable at no additional cost.

3084. Straight to Zero: Why Linearly Decaying the Learning Rate to Zero Works Best for LLMs

链接：https://iclr.cc/virtual/2025/poster/28738 abstract： LLMs are commonly trained with a learning rate (LR) warmup, followed by cosine decay to 10% of the maximum (10x decay). In a large-scale empirical study, we show that under an optimal peak LR, a simple linear decay-to-zero (D2Z) schedule consistently outperforms other schedules when training at compute-optimal dataset sizes. D2Z is superior across a range of model sizes, batch sizes, datasets, and vocabularies. Benefits increase as dataset size increases. Leveraging a novel interpretation of AdamW as an exponential moving average of weight updates, we show how linear D2Z optimally balances the demands of early training (moving away from initial conditions) and late training (averaging over more updates in order to mitigate gradient noise). In experiments, a 610M-parameter model trained for 80 tokens-per-parameter (TPP) using D2Z achieves lower loss than when trained for 200 TPP using 10x decay, corresponding to an astonishing 60% compute savings. Models such as Llama2-7B, trained for 286 TPP with 10x decay, could likely have saved a majority of compute by training with D2Z.

3085. BRAID: Input-driven Nonlinear Dynamical Modeling of Neural-Behavioral Data

链接：https://iclr.cc/virtual/2025/poster/31051 abstract： Neural populations exhibit complex recurrent structures that drive behavior, while continuously receiving and integrating external inputs from sensory stimuli, upstream regions, and neurostimulation. However, neural populations are often modeled as autonomous dynamical systems, with little consideration given to the influence of external inputs that shape the population activity and behavioral outcomes. Here, we introduce BRAID, a deep learning framework that models nonlinear neural dynamics underlying behavior while explicitly incorporating any measured external inputs. Our method disentangles intrinsic recurrent neural population dynamics from the effects of inputs by including a forecasting objective within input-driven recurrent neural networks. BRAID further prioritizes the learning of intrinsic dynamics that are related to a behavior of interest by using a multi-stage optimization scheme. We validate BRAID with nonlinear simulations, showing that it can accurately learn the intrinsic dynamics shared between neural and behavioral modalities. We then apply

BRAID to motor cortical activity recorded during a motor task and demonstrate that our method more accurately fits the neural-behavioral data by incorporating measured sensory stimuli into the model and improves the forecasting of neural-behavioral data compared with various baseline methods, whether input-driven or not.

3086. Learning Clustering-based Prototypes for Compositional Zero-Shot Learning

链接：https://iclr.cc/virtual/2025/poster/28941 abstract： Learning primitive (i.e., attribute and object) concepts from seen compositions is the primary challenge of Compositional Zero-Shot Learning (CZSL). Existing CZSL solutions typically rely on oversimplified data assumptions, e.g., modeling each primitive with a single centroid primitive presentation, ignoring the natural diversities of the attribute (resp. object) when coupled with different objects (resp. attribute). In this work, we develop ClusPro, a robust clustering-based prototype mining framework for CZSL that defines the conceptual boundaries of primitives through a set of diversified prototypes. Specifically, ClusPro conducts within-primitive clustering on the embedding space for automatically discovering and dynamically updating prototypes. To learn high-quality embeddings for discriminative prototype construction, ClusPro repaints a well-structured and independent primitive embedding space, ensuring intra-primitive separation and inter-primitive decorrelation through prototype-based contrastive learning and decorrelation learning. Moreover, ClusPro effectively performs prototype clustering in a non-parametric fashion without the introduction of additional learnable parameters or computational budget during testing. Experiments on three benchmarks demonstrate ClusPro outperforms various top-leading CZSL solutions under both closed-world and open-world settings. Our code is available at CLUSPRO.

3087. DenseGrounding: Improving Dense Language-Vision Semantics for Ego-centric 3D Visual Grounding

链接：https://iclr.cc/virtual/2025/poster/28704 abstract： Enabling intelligent agents to comprehend and interact with 3D environments through natural language is crucial for advancing robotics and human-computer interaction. A fundamental task in this field is ego-centric 3D visual grounding, where agents locate target objects in real-world 3D spaces based on verbal descriptions. However, this task faces two significant challenges: (1) loss of fine-grained visual semantics due to sparse fusion of point clouds with ego-centric multi-view images, (2) limited textual semantic context due to arbitrary language descriptions. We propose DenseGrounding, a novel approach designed to address these issues by enhancing both visual and textual semantics. For visual features, we introduce the Hierarchical Scene Semantic Enhancer, which retains dense semantics by capturing fine-grained global scene features and facilitating cross-modal alignment. For text descriptions, we propose a Language Semantic Enhancer that leverage large language models to provide rich context and diverse language descriptions with additional context during model training. Extensive experiments show that DenseGrounding significantly outperforms existing methods in overall accuracy, achieving improvements of 5.81% and 7.56% when trained on the comprehensive full training dataset and smaller mini subset, respectively, further advancing the SOTA in ego-centric 3D visual grounding. Our method also achieves 1st place and receives Innovation Award in the 2024 Autonomous Grand Challenge Multi-view 3D Visual Grounding Track, validating its effectiveness and robustness.

3088. Both Ears Wide Open: Towards Language-Driven Spatial Audio Generation

链接：https://iclr.cc/virtual/2025/poster/28250 abstract： Recently, diffusion models have achieved great success in mono-channel audio generation.However, when it comes to stereo audio generation, the soundscapes often have a complex scene of multiple objects and directions.Controlling stereo audio with spatial contexts remains challenging due to high data costs and unstable generative models. To the best of our knowledge, this work represents the first attempt to address these issues.We first construct a large-scale, simulation-based, and GPT-assisted dataset, BEWO-1M, with abundant soundscapes and descriptions even including moving and multiple sources.Beyond text modality, we have also acquired a set of images and rationally paired stereo audios through retrieval to advance multimodal generation. Existing audio generation models tend to generate rather random and indistinct spatial audio. To provide accurate guidance for Latent Diffusion Models, we introduce the SpatialSonic model utilizing spatial-aware encoders and azimuth state matrices to reveal reasonable spatial guidance. By leveraging spatial guidance, our model not only achieves the objective of generating immersive and controllable spatial audio from text but also extends to other modalities as the pioneer attempt.Finally, under fair settings, we conduct subjective and objective evaluations on simulated and real-world data to compare our approach with prevailing methods. The results demonstrate the effectiveness of our method, highlighting its capability to generate spatial audio that adheres to physical rules.

3089. CR2PQ: Continuous Relative Rotary Positional Query for Dense Visual Representation Learning

链接：https://iclr.cc/virtual/2025/poster/31058 abstract： Dense visual contrastive learning (DRL) shows promise for learning localized information in dense prediction tasks, but struggles with establishing pixel/patch correspondence across different views (cross-contrasting). Existing methods primarily rely on self-contrasting the same view with variations, limiting input variance and hindering downstream performance. This paper delves into the mechanisms of self-contrasting and cross-contrasting, identifying the crux of the issue: transforming discrete positional embeddings to continuous representations. To address the correspondence problem, we propose a Continuous Relative Rotary Positional Query ({\mname}), enabling patch-

level representation learning. Our extensive experiments on standard datasets demonstrate state-of-the-art (SOTA) results. Compared to the previous SOTA method (PQCL), our approach achieves significant improvements on COCO: with 300 epochs of pretraining, {\mname} obtains \textbf{3.4\%} mAP^{bb} and \textbf{2.1\%} mAP^{mk} improvements for detection and segmentation tasks, respectively. Furthermore, {\mname} exhibits faster convergence, achieving \textbf{10.4\%} mAP^{bb} and \textbf{7.9\%} mAP^{mk} improvements over SOTA with just 40 epochs of pretraining.

3090. Benign Overfitting in Out-of-Distribution Generalization of Linear Models

链接：https://iclr.cc/virtual/2025/poster/30869 abstract： Benign overfitting refers to the phenomenon where an over-parameterized model fits the training data perfectly, including noise in the data, but still generalizes well to the unseen test data. While prior work provides some theoretical understanding of this phenomenon under the in-distribution setup, modern machine learning often operates in a more challenging Out-of-Distribution (OOD) regime, where the target (test) distribution can be rather different from the source (training) distribution. In this work, we take an initial step towards understanding benign overfitting in the OOD regime by focusing on the basic setup of over-parameterized linear models under covariate shift. We provide non-asymptotic guarantees proving that benign overfitting occurs in standard ridge regression, even under the OOD regime when the target covariance satisfies certain structural conditions. We identify several vital quantities relating to source and target covariance, which govern the performance of OOD generalization. Our result is sharp, which provably recovers prior in-distribution benign overfitting guarantee (Tsigler & Bartlett, 2023), as well as under-parameterized OOD guarantee (Ge et al., 2024) when specializing to each setup. Moreover, we also present theoretical results for a more general family of target covariance matrix, where standard ridge regression only achieves a slow statistical rate of $\mathcal{O}(1/\sqrt{n})$ for the excess risk, while Principal Component Regression (PCR) is guaranteed to achieve the fast rate $\mathcal{O}(1/n)$, where n is the number of samples.

3091. Captured by Captions: On Memorization and its Mitigation in CLIP Models

链接：https://iclr.cc/virtual/2025/poster/30943 abstract： Multi-modal models, such as CLIP, have demonstrated strong performance in aligning visual and textual representations, excelling in tasks like image retrieval and zero-shot classification. Despite this success, the mechanisms by which these models utilize training data, particularly the role of memorization, remain unclear. In uni-modal models, both supervised and self-supervised, memorization has been shown to be essential for generalization. However, it is not well understood how these findings would apply to CLIP, which incorporates elements from both supervised learning via captions that provide a supervisory signal similar to labels, and from self-supervised learning via the contrastive objective.To bridge this gap in understanding, we propose a formal definition of memorization in CLIP (CLIPMem) and use it to quantify memorization in CLIP models. Our results indicate that CLIP’s memorization behavior falls between the supervised and self-supervised paradigms, with "mis-captioned" samples exhibiting highest levels of memorization. Additionally, we find that the text encoder contributes more to memorization than the image encoder, suggesting that mitigation strategies should focus on the text domain. Building on these insights, we propose multiple strategies to reduce memorization while at the same time improving utility---something that had not been shown before for traditional learning paradigms where reducing memorization typically results in utility decrease.

3092. Ready-to-React: Online Reaction Policy for Two-Character Interaction Generation

链接：https://iclr.cc/virtual/2025/poster/28451 abstract： This paper addresses the task of generating two-character online interactions. Previously, two main settings existed for two-character interaction generation: (1) generating one's motions based on the counterpart's complete motion sequence, and (2) jointly generating two-character motions based on specific conditions. We argue that these settings fail to model the process of real-life two-character interactions, where humans will react to their counterparts in real time and act as independent individuals. In contrast, we propose an online reaction policy, called Ready-to-React, to generate the next character pose based on past observed motions. Each character has its own reaction policy as its ``brain'', enabling them to interact like real humans in a streaming manner. Our policy is implemented by incorporating a diffusion head into an auto-regressive model, which can dynamically respond to the counterpart's motions while effectively mitigating the error accumulation throughout the generation process. We conduct comprehensive experiments using the challenging boxing task. Experimental results demonstrate that our method outperforms existing baselines and can generate extended motion sequences. Additionally, we show that our approach can be controlled by sparse signals, making it well-suited for VR and other online interactive environments. Code and data will be made publicly available.

3093. Understanding Fairness Surrogate Functions in Algorithmic Fairness

链接：https://iclr.cc/virtual/2025/poster/31494 abstract： It has been observed that machine learning algorithms exhibit biased predictions against certain population groups. To mitigate such bias while achieving comparable accuracy, a promising approach is to introduce surrogate functions of the concerned fairness definition and solve a constrained optimization problem. However, it is intriguing in previous work that such fairness surrogate functions may yield unfair results and high instability. In this work, in order to deeply understand them, taking a widely used fairness definition—demographic parity as an example, we show

that there is a surrogate-fairness gap between the fairness definition and the fairness surrogate function. Also, the theoretical analysis and experimental results about the “gap” motivate us that the fairness and stability will be affected by the points far from the decision boundary, which is the large margin points issue investigated in this paper. To address it, we propose the general sigmoid surrogate to simultaneously reduce both the surrogate-fairness gap and the variance, and offer a rigorous fairness and stability upper bound. Interestingly, the theory also provides insights into two important issues that deal with the large margin points as well as obtaining a more balanced dataset are beneficial to fairness and stability. Furthermore, we elaborate a novel and general algorithm called Balanced Surrogate, which iteratively reduces the “gap” to mitigate unfairness. Finally, we provide empirical evidence showing that our methods consistently improve fairness and stability while maintaining accuracy comparable to the baselines in three real-world datasets.

3094. Neural Eulerian Scene Flow Fields

链接：https://iclr.cc/virtual/2025/poster/31266 abstract： We reframe scene flow as the task of estimating a continuous space-time ordinary differential equation (ODE) that describes motion for an entire observation sequence, represented with a neural prior. Our method, EulerFlow, optimizes this neural prior estimate against several multi-observation reconstruction objectives, enabling high quality scene flow estimation via self-supervision on real-world data. EulerFlow works out-of-the-box without tuning across multiple domains, including large-scale autonomous driving scenes and dynamic tabletop settings. Remarkably, EulerFlow produces high quality flow estimates on small, fast moving objects like birds and tennis balls, and exhibits emergent 3D point tracking behavior by solving its estimated ODE over long-time horizons. On the Argoverse 2 2024 Scene Flow Challenge, EulerFlow outperforms all prior art, surpassing the next-best unsupervised method by more than 2.5 times, and even exceeding the next-best supervised method by over 10%. See https://vedder.io/eulerflow for interactive visuals.

3095. Eagle: Exploring The Design Space for Multimodal LLMs with Mixture of Encoders

链接：https://iclr.cc/virtual/2025/poster/29276 abstract： The ability to accurately interpret complex visual information is a crucial topic of multimodal large language models (MLLMs). Recent work indicates that enhanced visual perception significantly reduces hallucinations and improves performance on resolution-sensitive tasks, such as optical character recognition and document analysis. A number of recent MLLMs achieve this goal using a mixture of vision encoders. Despite their success, there is a lack of systematic comparisons and detailed ablation studies addressing critical aspects, such as expert selection and the integration of multiple vision experts. This study provides an extensive exploration of the design space for MLLMs using a mixture of vision encoders and resolutions. Our findings reveal several underlying principles common to various existing strategies, leading to a streamlined yet effective design approach. We discover that simply concatenating visual tokens from a set of complementary vision encoders is as effective as more complex mixing architectures or strategies. We additionally introduce Pre-Alignment to bridge the gap between vision-focused encoders and language tokens, enhancing model coherence. The resulting family of MLLMs, Eagle, surpasses other leading open-source models on major MLLM benchmarks.

3096. Calibrating LLMs with Information-Theoretic Evidential Deep Learning

链接：https://iclr.cc/virtual/2025/poster/29255 abstract： Fine-tuned large language models (LLMs) often exhibit overconfidence, particularly when trained on small datasets, resulting in poor calibration and inaccurate uncertainty estimates. Evidential Deep Learning (EDL), an uncertainty-aware approach, enables uncertainty estimation in a single forward pass, making it a promising method for calibrating fine-tuned LLMs. However, despite its computational efficiency, EDL is prone to overfitting, as its training objective can result in overly concentrated probability distributions. To mitigate this, we propose regularizing EDL by incorporating an information bottleneck (IB). Our approach IB-EDL suppresses spurious information in the evidence generated by the model and encourages truly predictive information to influence both the predictions and uncertainty estimates. Extensive experiments across various fine-tuned LLMs and tasks demonstrate that IB-EDL outperforms both existing EDL and non-EDL approaches. By improving the trustworthiness of LLMs, IB-EDL facilitates their broader adoption in domains requiring high levels of confidence calibration.

3097. Diffusion Actor-Critic: Formulating Constrained Policy Iteration as Diffusion Noise Regression for Offline Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28510 abstract： In offline reinforcement learning, it is necessary to manage out-of-distribution actions to prevent overestimation of value functions. One class of methods, the policy-regularized method, addresses this problem by constraining the target policy to stay close to the behavior policy. Although several approaches suggest representing the behavior policy as an expressive diffusion model to boost performance, it remains unclear how to regularize the target policy given a diffusion-modeled behavior sampler. In this paper, we propose Diffusion Actor-Critic (DAC) that formulates the Kullback-Leibler (KL) constraint policy iteration as a diffusion noise regression problem, enabling direct representation of target policies as diffusion models. Our approach follows the actor-critic learning paradigm in which we alternatively train a diffusion-modeled target policy and a critic network. The actor training loss includes a soft Q-guidance term from the Q-gradient. The soft Q-guidance is based on the theoretical solution of the KL constraint policy iteration, which prevents the learned policy from taking out-of-distribution actions. We demonstrate that such diffusion-based policy constraint, along with the coupling of the lower confidence bound of the Q-ensemble as value targets, not only preserves the multi-modality of target policies, but also contributes to stable convergence and strong performance in DAC. Our approach is evaluated on D4RL

benchmarks and outperforms the state-of-the-art in nearly all environments.

3098. TIGeR: Unifying Text-to-Image Generation and Retrieval with Large Multimodal Models

链接：https://iclr.cc/virtual/2025/poster/28444 abstract： How humans can effectively and efficiently acquire images has always been a perennial question. A classic solution is text-to-image retrieval from an existing database; however, the limited database typically lacks creativity. By contrast, recent breakthroughs in text-to-image generation have made it possible to produce attractive and counterfactual visual content, but it faces challenges in synthesizing knowledge-intensive images. In this work, we rethink the relationship between text-to-image generation and retrieval, proposing a unified framework for both tasks with one single Large Multimodal Model (LMM). Specifically, we first explore the intrinsic discriminative abilities of LMMs and introduce an efficient generative retrieval method for text-to-image retrieval in a training-free manner. Subsequently, we unify generation and retrieval autoregressively and propose an autonomous decision mechanism to choose the best-matched one between generated and retrieved images as the response to the text prompt. To standardize the evaluation of unified text-to-image generation and retrieval, we construct TIGeR-Bench, a benchmark spanning both creative and knowledge-intensive domains. Extensive experiments on TIGeR-Bench and two retrieval benchmarks, i.e., Flickr30K and MS-COCO, demonstrate the superiority of our proposed framework.

3099. Multi-modal Agent Tuning: Building a VLM-Driven Agent for Efficient Tool Usage

链接：https://iclr.cc/virtual/2025/poster/31249 abstract： The advancement of large language models (LLMs) prompts the development of multi-modal agents, which are used as a controller to call external tools, providing a feasible way to solve practical tasks. In this paper, we propose a multi-modal agent tuning method that automatically generates multi-modal tool-usage data and tunes a vision-language model (VLM) as the controller for powerful tool-usage reasoning. To preserve the data quality, we prompt the GPT-4o mini model to generate queries, files, and trajectories, followed by query-file and trajectory verifiers. Based on the data synthesis pipeline, we collect the MM-Traj dataset that contains 20K tasks with trajectories of tool usage. Then, we develop the T3-Agent via Trajectory Tuning on VLMs for Tool usage using MM-Traj. Evaluations on the GTA and GAIA benchmarks show that the T3-Agent consistently achieves improvements on two popular VLMs: MiniCPM-V-8.5B and Qwen2-VL-7B, which outperforms untrained VLMs by 20%, showing the effectiveness of the proposed data synthesis pipeline, leading to high-quality data for tool-usage capabilities.

3100. Effective and Efficient Time-Varying Counterfactual Prediction with State-Space Models

链接：https://iclr.cc/virtual/2025/poster/27707 abstract： Time-varying counterfactual prediction (TCP) from observational data supports the answer of when and how to assign multiple sequential treatments, yielding importance in various applications. Despite the progress achieved by recent advances, e.g., LSTM or Transformer based causal approaches, their capability of capturing interactions in long sequences remains to be improved in both prediction performance and running efficiency. In parallel with the development of TCP, the success of the state-space models (SSMs) has achieved remarkable progress toward long-sequence modeling with saved running time. Consequently, studying how Mamba simultaneously benefits the effectiveness and efficiency of TCP becomes a compelling research direction. In this paper, we propose to exploit advantages of the SSMs to tackle the TCP task, by introducing a counterfactual Mamba model with Covariate-based Decorrelation towards Selective Parameters (Mamba-CDSP). Motivated by the over-balancing problem in TCP of the direct covariate balancing methods, we propose to de-correlate between the current treatment and the representation of historical covariates, treatments, and outcomes, which can mitigate the confounding bias while preserve more covariate information. In addition, we show that the overall de-correlation in TCP is equivalent to regularizing the selective parameters of Mamba over each time step, which leads our approach to be effective and lightweight. We conducted extensive experiments on both synthetic and real-world datasets, demonstrating that Mamba-CDSP not only outperforms baselines by a large margin, but also exhibits prominent running efficiency.

3101. Rethinking Neural Multi-Objective Combinatorial Optimization via Neat Weight Embedding

链接：https://iclr.cc/virtual/2025/poster/30285 abstract： Recent decomposition-based neural multi-objective combinatorial optimization (MOCO) methods struggle to achieve desirable performance. Even equipped with complex learning techniques, they often suffer from significant optimality gaps in weight-specific subproblems. To address this challenge, we propose a neat weight embedding method to learn weight-specific representations, which captures weight-instance interaction for the subproblems and was overlooked by most current methods. We demonstrate the potentials of our method in two instantiations. First, we introduce a succinct addition model to learn weight-specific node embeddings, which surpassed most existing neural methods. Second, we design an enhanced conditional attention model to simultaneously learn the weight embedding and node embeddings, which yielded new state-of-the-art performance. Experimental results on classic MOCO problems verified the superiority of our method. Remarkably, our method also exhibits favorable generalization performance across problem sizes, even outperforming the neural method specialized for boosting size generalization.

3102. Neural Multi-Objective Combinatorial Optimization via Graph-Image Multimodal Fusion

链接：https://iclr.cc/virtual/2025/poster/30989 abstract： Existing neural multi-objective combinatorial optimization (MOCO) methods still exhibit an optimality gap since they fail to fully exploit the intrinsic features of problem instances. A significant factor contributing to this shortfall is their reliance solely on graph-modal information. To overcome this, we propose a novel graph-image multimodal fusion (GIMF) framework that enhances neural MOCO methods by integrating graph and image information of the problem instances. Our GIMF framework comprises three key components: (1) a constructed coordinate image to better represent the spatial structure of the problem instance, (2) a problem-size adaptive resolution strategy during the image construction process to improve the cross-size generalization of the model, and (3) a multimodal fusion mechanism with modality-specific bottlenecks to efficiently couple graph and image information. We demonstrate the versatility of our GIMF by implementing it with two state-of-the-art neural MOCO backbones. Experimental results on classic MOCO problems show that our GIMF significantly outperforms state-of-the-art neural MOCO methods and exhibits superior generalization capability.

3103. Adaptive Deployment of Untrusted LLMs Reduces Distributed Threats

链接：https://iclr.cc/virtual/2025/poster/28567 abstract： As large language models (LLMs) grow more powerful, they also become more difficult to trust. They could be either aligned with human intentions, or exhibit "subversive misalignment" -- introducing subtle errors that bypass safety checks. Although individual errors may not immediately cause harm, each increases the risk of an eventual safety failure. With this uncertainty, model deployment often grapples with the tradeoff between ensuring safety and harnessing the capabilities of untrusted models. In this work, we introduce the ``Diffuse Risk Management'' problem, aiming to balance the average-case safety and usefulness in the deployment of untrusted models over a large sequence of tasks. We approach this problem by developing a two-level framework: the single-task level (micro-protocol) and the whole-scenario level (macro-protocol). At the single-task level, we develop various \textit{micro}-protocols that use a less capable, but extensively tested (trusted) model to harness and monitor the untrusted model. At the whole-scenario level, we find an optimal \textit{macro}-protocol that uses an adaptive estimate of the untrusted model's risk to choose between micro-protocols. To evaluate the robustness of our method, we follow \textit{control evaluations} in a code generation testbed, which involves a red team attempting to generate subtly backdoored code with an LLM whose deployment is safeguarded by a blue team. Experiment results show that our approach retains 99.6\% usefulness of the untrusted model while ensuring near-perfect safety, significantly outperforming existing deployment methods. Our approach also demonstrates robustness when the trusted and untrusted models have a large capability gap. Our findings demonstrate the promise of managing diffuse risks in the deployment of increasingly capable but untrusted LLMs.

3104. Language Models Learn to Mislead Humans via RLHF

链接：https://iclr.cc/virtual/2025/poster/27791 abstract： Language models (LMs) can produce errors that are hard to detect for humans, especially when the task is complex.RLHF, the most popular post-training method, may exacerbate this problem: to achieve higher rewards, LMs might get better at convincing humans that they are right even when they are wrong. We study this phenomenon under a standard RLHF pipeline, calling it ``U-Sophistry'' since it is \textbf{U}nintended by model developers. Specifically, we ask time-constrained (e.g., 3-10 minutes) human subjects to evaluate the correctness of model outputs and calculate humans' accuracy against gold labels. On a question-answering task (QuALITY) and programming task (APPS), RLHF makes LMs better at convincing our subjects but not at completing the task correctly. RLHF also makes the model harder to evaluate: our subjects' false positive rate increases by 24.1% on QuALITY and 18.3% on APPS.Finally, we show that probing, a state-of-the-art approach for detecting \textbf{I}ntended Sophistry (e.g.~backdoored LMs), does not generalize to U-Sophistry. Our results highlight an important failure mode of RLHF and call for more research in assisting humans to align them.

3105. Adversarial Generative Flow Network for Solving Vehicle Routing Problems

链接：https://iclr.cc/virtual/2025/poster/28069 abstract： Recent research into solving vehicle routing problems (VRPs) has gained significant traction, particularly through the application of deep (reinforcement) learning for end-to-end solution construction. However, many current construction-based neural solvers predominantly utilize Transformer architectures, which can face scalability challenges and struggle to produce diverse solutions. To address these limitations, we introduce a novel framework beyond Transformer-based approaches, i.e., Adversarial Generative Flow Networks (AGFN). This framework integrates the generative flow network (GFlowNet)—a probabilistic model inherently adept at generating diverse solutions (routes)—with a complementary model for discriminating (or evaluating) the solutions. These models are trained alternately in an adversarial manner to improve the overall solution quality, followed by a proposed hybrid decoding method to construct the solution. We apply the AGFN framework to solve the capacitated vehicle routing problem (CVRP) and travelling salesman problem (TSP), and our experimental results demonstrate that AGFN surpasses the popular construction-based neural solvers, showcasing strong generalization capabilities on synthetic and real-world benchmark instances.

3106. DataMan: Data Manager for Pre-training Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28930 abstract： The performance emergence of large language models (LLMs) driven

by data scaling laws makes the selection of pre-training data increasingly important. However, existing methods rely on limited heuristics and human intuition, lacking comprehensive and clear guidelines.To address this, we are inspired by ``reverse thinking'' -- prompting LLMs to self-identify which criteria benefit its performance. As its pre-training capabilities are related to perplexity (PPL), we derive 14 quality criteria from the causes of text perplexity anomalies and introduce 15 common application domains to support domain mixing.In this paper, we train a Data Manager (DataMan) to learn quality ratings and domain recognition from pointwise rating, and use it to annotate a 447B token pre-training corpus with 14 quality ratings and domain type.Our experiments validate our approach, using DataMan to select 30B tokens to train a 1.3B-parameter language model, demonstrating significant improvements in in-context learning (ICL), perplexity, and instruction-following ability over the state-of-the-art baseline. The best-performing model, based on the Overall Score l=5 surpasses a model trained with 50% more data using uniform sampling. We continue pre-training with high-rated, domain-specific data annotated by DataMan to enhance domain-specific ICL performance and thus verify DataMan's domain mixing ability. Our findings emphasize the importance of quality ranking, the complementary nature of quality criteria, and their low correlation with perplexity, analyzing misalignment between PPL and ICL performance. We also thoroughly analyzed our pre-training dataset, examining its composition, the distribution of quality ratings, and the original document sources.

3107. Graph Assisted Offline-Online Deep Reinforcement Learning for Dynamic Workflow Scheduling

链接：https://iclr.cc/virtual/2025/poster/31012 abstract： Dynamic workflow scheduling (DWS) in cloud computing presents substantial challenges due to heterogeneous machine configurations, unpredictable workflow arrivals/patterns, and constantly evolving environments. However, existing research often assumes homogeneous setups and static conditions, limiting flexibility and adaptability in real-world scenarios. In this paper, we propose a novel Graph assisted Offline-Online Deep Reinforcement Learning (GOODRL) approach to building an effective and efficient scheduling agent for DWS. Our approach features three key innovations: (1) a task-specific graph representation and a Graph Attention Actor Network that enable the agent to dynamically assign focused tasks to heterogeneous machines while explicitly considering the future impact of each machine on these tasks; (2) a system-oriented graph representation and a Graph Attention Critic Network that facilitate efficient processing of new information and understanding its impact on the current state, crucial for managing unpredictable workflow arrivals/patterns in real-time; and (3) an offline-online method that utilizes imitation learning for effective offline training and applies gradient control and decoupled high-frequency critic training techniques during online learning to sustain the agent’s robust performance in rapidly changing environments. Experimental results demonstrate that GOODRL significantly outperforms several state-of-the-art algorithms, achieving substantially lower mean flowtime and high adaptability in various online and offline scenarios.

3108. RNNs are not Transformers (Yet): The Key Bottleneck on In-Context Retrieval

链接：https://iclr.cc/virtual/2025/poster/28783 abstract： This paper investigates the gap in representation powers of Transformers and Recurrent Neural Networks (RNNs), which are more memory efficient than Transformers. We aim to understand whether RNNs can match the performance of Transformers, particularly when enhanced with Chain-of-Thought (CoT) prompting. Our theoretical analysis reveals that CoT improves RNNs but is insufficient to close the gap with Transformers. A key bottleneck lies in the inability of RNNs to perfectly retrieve information from the context, even with CoT: for several tasks that explicitly or implicitly require this capability, such as associative recall and determining if a graph is a tree, we prove that RNNs are not expressive enough to solve the tasks while Transformers can solve them with ease.Conversely, we prove that adopting techniques to enhance the in-context retrieval capability of RNNs, including Retrieval-Augmented Generation (RAG) and adding a single Transformer layer, can elevate RNNs to be capable of solving all polynomial-time solvable problems with CoT, hence closing the representation gap with Transformers. We validate our theory on synthetic and natural language experiments.

3109. Efficient Action-Constrained Reinforcement Learning via Acceptance-Rejection Method and Augmented MDPs

链接：https://iclr.cc/virtual/2025/poster/30623 abstract： Action-constrained reinforcement learning (ACRL) is a generic framework for learning control policies with zero action constraint violation, which is required by various safety-critical and resource-constrained applications. The existing ACRL methods can typically achieve favorable constraint satisfaction but at the cost of either high computational burden incurred by the quadratic programs (QP) or increased architectural complexity due to the use of sophisticated generative models. In this paper, we propose a generic and computationally efficient framework that can adapt a standard unconstrained RL method to ACRL through two modifications: (i) To enforce the action constraints, we leverage the classic acceptance-rejection method, where we treat the unconstrained policy as the proposal distribution and derive a modified policy with feasible actions. (ii) To improve the acceptance rate of the proposal distribution, we construct an augmented two-objective Markov decision process (MDP), which include additional self-loop state transitions and a penalty signal for the rejected actions. This augmented MDP incentives the learned policy to stay close to the feasible action sets. Through extensive experiments in both robot control and resource allocation domains, we demonstrate that the proposed framework enjoys faster training progress, better constraint satisfaction, and a lower action inference time simultaneously than the state-of-the-art ACRL methods. We have made the source code publicly available to encourage further research in this direction.

3110. QMP: Q-switch Mixture of Policies for Multi-Task Behavior Sharing

链接：https://iclr.cc/virtual/2025/poster/29165 abstract： Multi-task reinforcement learning (MTRL) aims to learn several tasks simultaneously for better sample efficiency than learning them separately. Traditional methods achieve this by sharing parameters or relabeling data between tasks. In this work, we introduce a new framework for sharing behavioral policies across tasks, which can be used in addition to existing MTRL methods. The key idea is to improve each task's off-policy data collection by employing behaviors from other task policies. Selectively sharing helpful behaviors acquired in one task to collect training data for another task can lead to higher-quality trajectories, leading to more sample-efficient MTRL. Thus, we introduce a simple and principled framework called Q-switch mixture of policies (QMP) that selectively shares behavior between different task policies by using the task's Q-function to evaluate and select useful shareable behaviors. We theoretically analyze how QMP improves the sample efficiency of the underlying RL algorithm. Our experiments show that QMP's behavioral policy sharing provides complementary gains over many popular MTRL algorithms and outperforms alternative ways to share behaviors in various manipulation, locomotion, and navigation environments. Videos are available at https://qmp-mtrl.github.io/.

3111. Flow Distillation Sampling: Regularizing 3D Gaussians with Pre-trained Matching Priors

链接：https://iclr.cc/virtual/2025/poster/30536 abstract： 3D Gaussian Splatting (3DGS) has achieved excellent rendering quality with fast training and rendering speed. However, its optimization process lacks explicit geometric constraints, leading to suboptimal geometric reconstruction in regions with sparse or no observational input views. In this work, we try to mitigate the issue by incorporating a pre-trained matching prior to the 3DGS optimization process. We introduce Flow Distillation Sampling (FDS), a technique that leverages pre-trained geometric knowledge to bolster the accuracy of the Gaussian radiance field. Our method employs a strategic sampling technique to target unobserved views adjacent to the input views, utilizing the optical flow calculated from the matching model (Prior Flow) to guide the flow analytically calculated from the 3DGS geometry (Radiance Flow). Comprehensive experiments in depth rendering, mesh reconstruction, and novel view synthesis showcase the significant advantages of FDS over state-of-the-art methods. Additionally, our interpretive experiments and analysis aim to shed light on the effects of FDS on geometric accuracy and rendering quality, potentially providing readers with insights into its performance.

3112. UniCoTT: A Unified Framework for Structural Chain-of-Thought Distillation

链接：https://iclr.cc/virtual/2025/poster/31068 abstract： Chains of thought (CoTs) have achieved success in enhancing the reasoning capabilities of large language models (LLMs), while their effectiveness is predominantly observed in LLMs. Existing solutions methods adopt distillation to inject chain-of-thought capabilities into small models (SLMs). However, they: (1) can not guarantee the rationality of the generated explanation due to hallucinations; (2) ignore diverse structures of CoT during knowledge transfer. In this paper, we propose a unified CoT distillation framework termed UniCoTT for considering diverse structural CoTs (\emph{i.e.}, chain, tree, and graph). UniCoTT contains two core strategies: iterative construction for structured CoTs and the structural constraint strategy. Specifically, UniCoTT prompts LLMs to iteratively produce accurate explanations with answers and unifies structured explanations as UniCoT which is seen as a bridge for knowledge transfer. Furthermore, UniCoTT utilizes the proposed unified supervised learning and structural consistency learning strategies to transfer knowledge of structured CoT to SLMs. Experimental results show that UniCoTT can significantly improve the performance of SLMs on multiple datasets across different NLP tasks. Our code is available at https://github.com/mengchuang123/UniCoTT.

3113. DistRL: An Asynchronous Distributed Reinforcement Learning Framework for On-Device Control Agent

链接：https://iclr.cc/virtual/2025/poster/30000 abstract： On-device control agents, especially on mobile devices, are responsible for operating mobile devices to fulfill users' requests, enabling seamless and intuitive interactions. Integrating Multimodal Large Language Models (MLLMs) into these agents enhances their ability to understand and execute complex commands, thereby improving user experience. However, fine-tuning MLLMs for on-device control presents significant challenges due to limited data availability and inefficient online training processes. This paper introduces DistRL, a novel framework designed to enhance the efficiency of online RL fine-tuning for mobile device control agents. DistRL employs centralized training and decentralized data acquisition to ensure efficient fine-tuning in the context of dynamic online interactions. Additionally, the framework is backed by our tailor-made RL algorithm, which effectively balances exploration with the prioritized utilization of collected data to ensure stable and robust training. Our experiments show that, on average, DistRL delivers a 3$\times$ improvement in training efficiency and enables training data collection 2.4$\times$ faster than the leading synchronous multi-machine methods. Notably, after training, DistRL achieves a 20\% relative improvement in success rate compared to state-of-the-art methods on general Android tasks from an open benchmark, significantly outperforming existing approaches while maintaining the same training time. These results validate DistRL as a scalable and efficient solution, offering substantial improvements in both training efficiency and agent performance for real-world, in-the-wild device control tasks.

3114. D_{2}O: Dynamic Discriminative Operations for Efficient Long-Context Inference of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30197 abstract： Efficient generative inference in Large Language Models (LLMs) is impeded by the growing memory demands of Key-Value (KV) cache, especially for longer sequences. Traditional KV Cache eviction strategies, which discard less critical KV-pairs based on attention scores, often degrade generation quality, leading to issues such as context loss or hallucinations. To address this, we introduce **D**ynamic **D**iscriminative **O**perations ($\mathbf{D_2 O}$), a novel method that optimizes KV cache size dynamically and discriminatively at two levels without fine-tuning, while preserving essential context. At **layer-level**, by observing the varying densities of attention weights between shallow and deep layers, we dynamically determine which layers should avoid excessive eviction via our proposed ***dynamic allocation strategy*** to minimize information loss. At **token-level**, for the eviction strategy in each layer, $\mathbf{D_2 O}$ innovatively incorporates a ***compensation mechanism*** that maintains a similarity threshold to re-discriminate the importance of currently discarded tokens, determining whether they should be recalled and merged with similar tokens. Extensive experiments on various benchmarks and LLM architectures have shown that $\mathbf{D_2 O}$ not only achieves significant memory savings and enhances inference throughput by more than 3$\times$ but also maintains high-quality long-text generation.

3115. Can Textual Gradient Work in Federated Learning?

链接：https://iclr.cc/virtual/2025/poster/30485 abstract： Recent studies highlight the promise of LLM-based prompt optimization, especially with TextGrad, which automates ``differentiation'' via texts and backpropagates textual feedback provided by LLMs. This approach facilitates training in various real-world applications that do not support numerical gradient propagation or loss calculation. It opens new avenues for optimization in decentralized, resource-constrained environments, suggesting that users of black-box LLMs (e.g., ChatGPT) could enhance components of LLM agentic systems (such as prompt optimization) through collaborative paradigms like federated learning (FL). In this paper, we systematically explore the potential and challenges of incorporating textual gradient into FL. Our contributions are fourfold.Firstly, we introduce a novel FL paradigm, Federated Textual Gradient (FedTextGrad), that allows FL clients to upload their locally optimized prompts derived from textual gradients, while the FL server aggregates the received prompts through text summarization. Unlike traditional FL frameworks, which are designed for numerical aggregation, FedTextGrad is specifically tailored for handling textual data, expanding the applicability of FL to a broader range of problems that lack well-defined numerical loss functions. Secondly, building on this design, we conduct extensive experiments to explore the feasibility of federated textual gradients. Our findings highlight the importance of properly tuning key factors (e.g., local steps) in FL training to effectively integrate textual gradients. Thirdly, we highlight a major challenge in federated textual gradient aggregation: retaining essential information from distributed prompt updates. Concatenation often produces prompts that exceed the LLM API's context window, while summarization can degrade performance by generating overly condensed or complex text that lacks key context. Last but not least, in response to this issue, we improve the vanilla variant of FedTextGrad by providing actionable guidance to the LLM when summarizing client prompts by leveraging the Uniform Information Density principle. Such a design reduces the complexity of the aggregated global prompt, thereby better incentivizing the LLM's reasoning ability. Through this principled study, we enable the adoption of textual gradients in FL for optimizing LLMs, identify important issues, and pinpoint future directions, thereby opening up a new research area that warrants further investigation.

3116. Unveiling the Secret Recipe: A Guide For Supervised Fine-Tuning Small LLMs

链接：https://iclr.cc/virtual/2025/poster/28940 abstract： The rise of large language models (LLMs) has created a significant disparity: industrial research labs with their computational resources, expert teams, and advanced infrastructures, can effectively fine-tune LLMs, while individual developers and small organizations face barriers due to limited resources to effectively explore the experiment space. In this paper, we aim to bridge this gap by presenting a comprehensive study on supervised fine-tuning of LLMs using instruction-tuning datasets spanning diverse knowledge domains and skills. We focus on small-sized LLMs (3B to 7B parameters) for their cost-efficiency and accessibility. We explore various training configurations and strategies across four open-source pre-trained models. We provide detailed documentation of these configurations, revealing findings that challenge several common training practices, including hyperparameter recommendations from TULU and phased training recommended by Orca. The code used for the experiments can be found here: https://github.com/instructlab/training.Key insights from our work include: (i) larger batch sizes paired with lower learning rates lead to improved model performance on benchmarks such as MMLU, MTBench, and Open LLM Leaderboard; (ii) early-stage training dynamics, such as lower gradient norms and higher loss values, are strong indicators of better final model performance, allowing for early termination of sub-optimal runs and significant computational savings; (iii) through a thorough exploration of hyperparameters like warmup steps and learning rate schedules, we provide guidance for practitioners and find that certain simplifications do not compromise performance; and (iv) we observe no significant difference in performance between phased (sequentially training on data divided into phases) and stacked (training on the entire dataset at once) strategies, but stacked training is simpler and more sample efficient. With these findings holding robustly across datasets as well as model families and sizes, we hope this study serves as a guide for practitioners fine-tuning small LLMs and promotes a more inclusive research environment for LLM development.

3117. Differentially Private Steering for Large Language Model Alignment

链接：https://iclr.cc/virtual/2025/poster/28523 abstract： Aligning Large Language Models (LLMs) with human values and away from undesirable behaviors (such as hallucination) has become increasingly important. Recently, steering LLMs towards a desired behavior via activation editing has emerged as an effective method to mitigate harmful generations at inference-time. Activation editing modifies LLM representations by preserving information from positive demonstrations (e.g., truthful) and minimising information from negative demonstrations (e.g., hallucinations). When these demonstrations come from a private

dataset, the aligned LLM may leak private information contained in those private samples. In this work, we present the first study of aligning LLM behavior with private datasets. Our work proposes the \textit{\underline{P}rivate \underline{S}teering for LLM \underline{A}lignment (PSA)} algorithm to edit LLM activations with differential privacy (DP) guarantees. We conduct extensive experiments on seven different benchmarks with open-source LLMs of different sizes (0.5B to 7B) and model families (LlaMa and Qwen). Our results show that PSA achieves DP guarantees for LLM alignment with minimal loss in performance, including alignment metrics, open-ended text generation quality, and general-purpose reasoning. We also develop the first Membership Inference Attack (MIA) for evaluating and auditing the empirical privacy for the problem of LLM steering via activation editing. Our attack is tailored for activation editing and relies solely on the generated texts without their associated probabilities. Our experiments support the theoretical guarantees by showing improved guarantees for our \textit{PSA} algorithm compared to several existing non-private techniques.

3118. Polynomial Composition Activations: Unleashing the Dynamics of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30507 abstract： Transformers have found extensive applications across various domains due to their powerful fitting capabilities. This success can be partially attributed to their inherent nonlinearity. Thus, in addition to the ReLU function employed in the original transformer architecture, researchers have explored alternative modules such as GeLU and SwishGLU to enhance nonlinearity and thereby augment representational capacity. In this paper, we propose a novel category of polynomial composition activations (PolyCom), designed to optimize the dynamics of transformers. Theoretically, we provide a comprehensive mathematical analysis of PolyCom, highlighting its enhanced expressivity and efficacy relative to other activation functions. Notably, we demonstrate that networks incorporating PolyCom achieve the optimal approximation rate, indicating that PolyCom networks require minimal parameters to approximate general smooth functions in Sobolev spaces. We conduct empirical experiments on the pre-training configurations of large language models (LLMs), including both dense and sparse architectures. By substituting conventional activation functions with PolyCom, we enable LLMs to capture higher-order interactions within the data, thus improving performance metrics in terms of accuracy and convergence rates. Extensive experimental results demonstrate the effectiveness of our method, showing substantial improvements over other activation functions. Code is available at https://github.com/BryceZhuo/PolyCom.

3119. MatryoshkaKV: Adaptive KV Compression via Trainable Orthogonal Projection

链接：https://iclr.cc/virtual/2025/poster/30571 abstract： KV cache has become a de facto technique for the inference of large language models (LLMs), where tensors of shape (layer number, head number, sequence length, feature dimension) are introduced to cache historical information for self-attention. As the size of the model and data grows, the KV cache can, yet, quickly become a bottleneck within the system in both storage and memory transfer.To address this, prior studies usually focus on the first three axes of the cache tensors for compression. This paper supplements them, focusing on the feature dimension axis, by utilizing low-rank projection matrices to transform the cache features into spaces with reduced dimensions. We begin by investigating the canonical orthogonal projection method for data compression through principal component analysis (PCA). We identify the drawback of PCA projection that model performance degrades rapidly under relatively low compression rates (less than 60%).This phenomenon is elucidated by insights derived from the principles of attention mechanisms.To bridge the gap, we propose to directly tune the orthogonal projection matrix on the continual pre-training or supervised fine-tuning datasets with an elaborate Matryoshka learning strategy.Thanks to such a strategy, we can adaptively search for the optimal compression rates for various layers and heads given varying compression budgets. Compared to Multi-head Latent Attention (MLA), our method can easily embrace pre-trained LLMs and hold a smooth tradeoff between performance and compression rate. We witness the high data efficiency of our training procedure and find that our method can sustain over 90\% performance with an average KV cache compression rate of 60% (and up to 75% in certain extreme scenarios) for popular LLMs like LLaMA2 and Mistral.

3120. Aligning Language Models with Demonstrated Feedback

链接：https://iclr.cc/virtual/2025/poster/31183 abstract： Language models are aligned to emulate the collective voice of many, resulting in outputs that align with no one in particular. Steering LLMs away from generic output is possible through supervised finetuning or RLHF, but requires prohibitively large datasets for new ad-hoc tasks. We argue that it is instead possible to align an LLM to a specific setting by leveraging a very small number (<10) of demonstrations as feedback. Our method, Demonstration ITerated Task Optimization (DITTO), directly aligns language model outputs to a user's demonstrated behaviors. Derived using ideas from online imitation learning, DITTO cheaply generates online comparison data by treating users' demonstrations as preferred over output from the LLM and its intermediate checkpoints. We evaluate DITTO's ability to learn fine-grained style and task alignment across domains such as news articles, emails, and blog posts. Additionally, we conduct a user study soliciting a range of demonstrations from participants ($N=16$). Across our benchmarks and user study, we find that win-rates for DITTO outperform few-shot prompting, supervised fine-tuning, and other self-play methods by an average of 19\% points. By using demonstrations as feedback directly, DITTO offers a novel method for effective customization of LLMs.

3121. AugKD: Ingenious Augmentations Empower Knowledge Distillation for Image Super-Resolution

链接：https://iclr.cc/virtual/2025/poster/30646 abstract： Knowledge distillation (KD) compresses deep neural networks by transferring task-related knowledge from cumbersome pre-trained teacher models to more compact student models. However, vanilla KD for image super-resolution (SR) networks yields only limited improvements due to the inherent nature of SR tasks, where the outputs of teacher models are noisy approximations of high-quality label images. In this work, we show that the potential of vanilla KD has been underestimated and demonstrate that the ingenious application of data augmentation methods can close the gap between it and more complex, well-designed methods. Unlike conventional training processes typically applying image augmentations simultaneously to both low-quality inputs and high-quality labels, we propose AugKD utilizing unpaired data augmentations to 1) generate auxiliary distillation samples and 2) impose label consistency regularization. Comprehensive experiments show that the AugKD significantly outperforms existing state-of-the-art KD methods across a range of SR tasks.

3122. Semantic Image Inversion and Editing using Rectified Stochastic Differential Equations

链接：https://iclr.cc/virtual/2025/poster/30203 abstract： Generative models transform random noise into images, while their inversion aims to reconstruct structured noise for recovery and editing.This paper addresses two key tasks: (i) inversion and (ii) editing of real images using stochastic equivalents of rectified flow models (e.g., Flux).While Diffusion Models (DMs) dominate the field of generative modeling for images, their inversion suffers from faithfulness and editability challenges due to nonlinear drift and diffusion.Existing DM inversion methods require costly training of additional parameters or test-time optimization of latent variables.Rectified Flows (RFs) offer a promising alternative to DMs, yet their inversion remains underexplored. We propose RF inversion using dynamic optimal control derived via a linear quadratic regulator, and prove that the resulting vector field is equivalent to a rectified stochastic differential equation. We further extend our framework to design a stochastic sampler for Flux.Our method achieves state-of-the-art performance in zero-shot inversion and editing, surpassing prior works in stroke-to-image synthesis and semantic image editing, with large-scale human evaluations confirming user preference.See our project page https://rf-inversion.github.io/ for code and demo.

3123. Uni^{2}Det: Unified and Universal Framework for Prompt-Guided Multi-dataset 3D Detection

链接：https://iclr.cc/virtual/2025/poster/30625 abstract： We present Uni^{2}Det, a brand new framework for unified and universal multi-dataset training on 3D detection, enabling robust performance across diverse domains and generalization to unseen domains. Due to substantial disparities in data distribution and variations in taxonomy across diverse domains, training such a detector by simply merging datasets poses a significant challenge. Motivated by this observation, we introduce multi-stage prompting modules for multi-dataset 3D detection, which leverages prompts based on the characteristics of corresponding datasets to mitigate existing differences. This elegant design facilitates seamless plug-and-play integration within various advanced 3D detection frameworks in a unified manner, while also allowing straightforward adaptation for universal applicability across datasets. Experiments are conducted across multiple dataset consolidation scenarios involving KITTI, Waymo, and nuScenes, demonstrating that our Uni^{2}Det outperforms existing methods by a large margin in multi-dataset training. Notably, results on zero-shot cross-dataset transfer validate the generalization capability of our proposed method. Our code is available at https://github.com/ThomasWangY/Uni2Det.

3124. Enhancing Learning with Label Differential Privacy by Vector Approximation

链接：https://iclr.cc/virtual/2025/poster/30142 abstract： Label differential privacy (DP) is a framework that protects the privacy of labels in training datasets, while the feature vectors are public. Existing approaches protect the privacy of labels by flipping them randomly, and then train a model to make the output approximate the privatized label. However, as the number of classes K increases, stronger randomization is needed, thus the performances of these methods become significantly worse. In this paper, we propose a vector approximation approach for learning with label local differential privacy, which is easy to implement and introduces little additional computational overhead. Instead of flipping each label into a single scalar, our method converts each label into a random vector with K components, whose expectations reflect class conditional probabilities. Intuitively, vector approximation retains more information than scalar labels. A brief theoretical analysis shows that the performance of our method only decays slightly with K. Finally, we conduct experiments on both synthesized and real datasets, which validate our theoretical analysis as well as the practical performance of our method.

3125. RB-Modulation: Training-Free Stylization using Reference-Based Modulation

链接：https://iclr.cc/virtual/2025/poster/29091 abstract： We propose Reference-Based Modulation (RB-Modulation), a new plug-and-play solution for training-free personalization of diffusion models.Existing training-free approaches exhibit difficulties in (a) style extraction from reference images in the absence of additional style or content text descriptions, (b) unwanted content leakage from reference style images, and (c) effective composition of style and content. RB-Modulation is built on a novel stochastic optimal controller where a style descriptor encodes the desired attributes through a terminal cost. The resulting drift not only overcomes the difficulties above, but also ensures high fidelity to the reference style and adheres to the given text

prompt. We also introduce a cross-attention-based feature aggregation scheme that allows RB-Modulation to decouple content and style from the reference image.With theoretical justification and empirical evidence, our test-time optimization framework demonstrates precise extraction and control of content and style in a training-free manner. Further, our method allows a seamless composition of content and style, which marks a departure from the dependency on external adapters or ControlNets. See project page: https://rb-modulation.github.io/ for code and further details.

3126. PAL: Sample-Efficient Personalized Reward Modeling for Pluralistic Alignment

链接：https://iclr.cc/virtual/2025/poster/31187 abstract： Foundation models trained on internet-scale data benefit from extensive alignment to human preferences before deployment. However, existing methods typically assume a homogeneous preference shared by all individuals, overlooking the diversity inherent in human values. In this work, we propose a general reward modeling framework for pluralistic alignment (PAL), which incorporates diverse preferences from the ground up. PAL has a modular design that leverages commonalities across users while catering to individual personalization, enabling efficient few-shot localization of preferences for new users. Extensive empirical evaluation demonstrates that PAL matches or outperforms state-of-the-art methods on both text-to-text and text-to-image tasks: on Reddit TL;DR Summary, PAL is 1.7% more accurate for seen users and 36% more accurate for unseen users compared to the previous best method, with 100× less parameters. On Pick-a-Pic v2, PAL is 2.5% more accurate than the best method with 156× fewer learned parameters. Finally, we provide theoretical analysis for generalization of rewards learned via PAL framework showcasing the reduction in number of samples needed per user.

3127. $InterLCM$: Low-Quality Images as Intermediate States of Latent Consistency Models for Effective Blind Face Restoration

链接：https://iclr.cc/virtual/2025/poster/28192 abstract： Diffusion priors have been used for blind face restoration (BFR) by fine-tuning diffusion models (DMs) on restoration datasets to recover low-quality images. However, the naive application of DMs presents several key limitations. (i) The diffusion prior has inferior semantic consistency (e.g., ID, structure and color.), increasing the difficulty of optimizing the BFR model;(ii) reliance on hundreds of denoising iterations, preventing the effective cooperation with perceptual losses, which is crucial for faithful restoration.Observing that the latent consistency model (LCM) learns consistency noise-to-data mappings on the ODE-trajectory and therefore shows more semantic consistency in the subject identity, structural information and color preservation, we propose $\textit{InterLCM}$ to leverage the LCM for its superior semantic consistency and efficiency to counter the above issues. Treating low-quality images as the intermediate state of LCM, $\textit{InterLCM}$ achieves a balance between fidelity and quality by starting from earlier LCM steps. LCM also allows the integration of perceptual loss during training, leading to improved restoration quality, particularly in real-world scenarios.To mitigate structural and semantic uncertainties, $\textit{InterLCM}$ incorporates a Visual Module to extract visual features and a Spatial Encoder to capture spatial details, enhancing the fidelity of restored images.Extensive experiments demonstrate that $\textit{InterLCM}$ outperforms existing approaches in both synthetic and real-world datasets while also achieving faster inference speed. Code and models will be publicly available.

3128. HelpSteer2-Preference: Complementing Ratings with Preferences

链接：https://iclr.cc/virtual/2025/poster/29917 abstract： Reward models are critical for aligning models to follow instructions, and are typically trained following one of two popular paradigms: Bradley-Terry style or Regression style. However, there is a lack of evidence that either approach is better than the other, when adequately matched for data. This is primarily because these approaches require data collected in different (but incompatible) formats, meaning that adequately matched data is not available in existing public datasets. To tackle this problem, we release preference annotations (designed for Bradley-Terry training) to complement existing ratings (designed for Regression style training) in the HelpSteer2 dataset. To improve data interpretability, preference annotations are accompanied with human-written justifications. Using this data, we conduct the first head-to-head comparison of Bradley-Terry and Regression models when adequately matched for data. Based on insights derived from such a comparison, we propose a novel approach to combine Bradley-Terry and Regression reward modeling. A Llama-3.1-70B-Instruct model tuned with this approach scores 94.1 on RewardBench, emerging top of more than 140 reward models as of 1 Oct 2024. This reward model can then be used with REINFORCE algorithm (RLHF) to align an Instruct model to reach 85.0 on Arena Hard, which is No. 1 as of 1 Oct 2024.We open-source this dataset (CC-BY-4.0 license) at https://huggingface.co/datasets/nvidia/HelpSteer2#preferences-new---1-oct-2024 and openly release the trained Reward and Instruct models at https://huggingface.co/nvidia/Llama-3.1-Nemotron-70B-Reward and https://huggingface.co/nvidia/Llama-3.1-Nemotron-70B-Instruct .

3129. Single Teacher, Multiple Perspectives: Teacher Knowledge Augmentation for Enhanced Knowledge Distillation

链接：https://iclr.cc/virtual/2025/poster/30441 abstract： Do diverse perspectives help students learn better? Multi-teacher knowledge distillation, which is a more effective technique than traditional single-teacher methods, supervises the student from different perspectives (i.e., teacher). While effective, multi-teacher, teacher ensemble, or teaching assistant-based approaches are computationally expensive and resource-intensive, as they require training multiple teacher networks. These concerns raise

a question: can we supervise the student with diverse perspectives using only a single teacher? We, as the pioneer, demonstrate TeKAP, a novel teacher knowledge augmentation technique that generates multiple synthetic teacher knowledge by perturbing the knowledge of a single pretrained teacher i.e., Teacher Knowledge Augmentation via Perturbation, at both the feature and logit levels. These multiple augmented teachers simulate an ensemble of models together. The student model is trained on both the actual and augmented teacher knowledge, benefiting from the diversity of an ensemble without the need to train multiple teachers. TeKAP significantly reduces training time and computational resources, making it feasible for large-scale applications and easily manageable. Experimental results demonstrate that our proposed method helps existing state-of-the-art knowledge distillation techniques achieve better performance, highlighting its potential as a cost-effective alternative. The source code can be found in the supplementary.

3130. Contextual Document Embeddings

链接：https://iclr.cc/virtual/2025/poster/29341 abstract： Dense document embeddings are central to neural retrieval. The dominant paradigm is to train and construct embeddings by running encoders directly on individual documents. In this work, we argue that these embeddings, while effective, are implicitly out-of-context for targeted use cases of retrieval, and that a contextualized document embedding should take into account both the document and neighboring documents in context - analogous to contextualized word embeddings. We propose two complementary methods for contextualized document embeddings: first, an alternative contrastive learning objective that explicitly incorporates the document neighbors into the intra-batch contextual loss; second, a new contextual architecture that explicitly encodes neighbor document information into the encoded representation. Results show that both methods achieve better performance than biencoders in several settings, with differences especially pronounced out-of-domain. We achieve state-of-the-art results on the MTEB benchmark with no hard negative mining, score distillation, dataset-specific instructions, intra-GPU example-sharing, or extremely large batch sizes. Our method can be applied to improve performance on any contrastive learning dataset and any biencoder.

3131. Matcha: Mitigating Graph Structure Shifts with Test-Time Adaptation

链接：https://iclr.cc/virtual/2025/poster/30388 abstract： Powerful as they are, graph neural networks (GNNs) are known to be vulnerable to distribution shifts. Recently, test-time adaptation (TTA) has attracted attention due to its ability to adapt a pre-trained model to a target domain, without re-accessing the source domain. However, existing TTA algorithms are primarily designed for attribute shifts in vision tasks, where samples are independent. These methods perform poorly on graph data that experience structure shifts, where node connectivity differs between source and target graphs. We attribute this performance gap to the distinct impact of node attribute shifts versus graph structure shifts: the latter significantly degrades the quality of node representations and blurs the boundaries between different node categories. To address structure shifts in graphs, we propose Matcha, an innovative framework designed for effective and efficient adaptation to structure shifts by adjusting the htop-aggregation parameters in GNNs. To enhance the representation quality, we design a prediction-informed clustering loss to encourage the formation of distinct clusters for different node categories. Additionally, Matcha seamlessly integrates with existing TTA algorithms, allowing it to handle attribute shifts effectively while improving overall performance under combined structure and attribute shifts. We validate the effectiveness of Matcha on both synthetic and real-world datasets, demonstrating its robustness across various combinations of structure and attribute shifts. Our codeis available at https://github.com/baowenxuan/Matcha.

3132. Video In-context Learning: Autoregressive Transformers are Zero-Shot Video Imitators

链接：https://iclr.cc/virtual/2025/poster/27826 abstract： People interact with the real-world largely dependent on visual signal, which are ubiquitous and illustrate detailed demonstrations. In this paper, we explore utilizing visual signals as a new interface for models to interact with the environment. Specifically, we choose videos as a representative visual signal. And by training autoregressive Transformers on video datasets in a self-supervised objective, we find that the model emerges a zero-shot capability to infer the semantics from a demonstration video, and imitate the semantics to an unseen scenario. This allows the models to perform unseen tasks by watching the demonstration video in an in-context manner, without further fine-tuning. To validate the imitation capacity, we design various evaluation metrics including both objective and subjective measures. The results show that our models can generate high-quality video clips that accurately align with the semantic guidance provided by the demonstration videos, and we also show that the imitation capacity follows the scaling law.

3133. FakeShield: Explainable Image Forgery Detection and Localization via Multi-modal Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28315 abstract： The rapid development of generative AI is a double-edged sword, which not only facilitates content creation but also makes image manipulation easier and more difficult to detect. Although current image forgery detection and localization (IFDL) methods are generally effective, they tend to face two challenges: \textbf{1)} black-box nature with unknown detection principle, \textbf{2)} limited generalization across diverse tampering methods (e.g., Photoshop, DeepFake, AIGC-Editing). To address these issues, we propose the explainable IFDL task and design FakeShield, a multi-modal framework capable of evaluating image authenticity, generating tampered region masks, and providing a judgment basis based on pixel-level and image-level tampering clues. Additionally, we leverage GPT-4o to enhance existing IFDL datasets, creating the Multi-Modal Tamper Description dataSet (MMTD-Set) for training FakeShield's tampering

analysis capabilities. Meanwhile, we incorporate a Domain Tag-guided Explainable Forgery Detection Module (DTE-FDM) and a Multi-modal Forgery Localization Module (MFLM) to address various types of tamper detection interpretation and achieve forgery localization guided by detailed textual descriptions. Extensive experiments demonstrate that FakeShield effectively detects and localizes various tampering techniques, offering an explainable and superior solution compared to previous IFDL methods. The code is available at https://github.com/zhipeixu/FakeShield.

3134. SecureGS: Boosting the Security and Fidelity of 3D Gaussian Splatting Steganography

链接：https://iclr.cc/virtual/2025/poster/30245 abstract： 3D Gaussian Splatting (3DGS) has emerged as a premier method for 3D representation due to its real-time rendering and high-quality outputs, underscoring the critical need to protect the privacy of 3D assets. Traditional NeRF steganography methods fail to address the explicit nature of 3DGS since its point cloud files are publicly accessible. Existing GS steganography solutions mitigate some issues but still struggle with reduced rendering fidelity, increased computational demands, and security flaws, especially in the security of the geometric structure of the visualized point cloud. To address these demands, we propose a \textbf{SecureGS}, a secure and efficient 3DGS steganography framework inspired by Scaffold-GS's anchor point design and neural decoding. SecureGS uses a hybrid decoupled Gaussian encryption mechanism to embed offsets, scales, rotations, and RGB attributes of the hidden 3D Gaussian points in anchor point features, retrievable only by authorized users through privacy-preserving neural networks. To further enhance security, we propose a density region-aware anchor growing and pruning strategy that adaptively locates optimal hiding regions without exposing hidden information. Extensive experiments show that SecureGS significantly surpasses existing GS steganography methods in rendering fidelity, speed, and security.

3135. Your Weak LLM is Secretly a Strong Teacher for Alignment

链接：https://iclr.cc/virtual/2025/poster/28137 abstract： The burgeoning capabilities of large language models (LLMs) have underscored the need for alignment to ensure these models act in accordance with human values and intentions. Existing alignment frameworks present constraints either in the form of expensive human effort or high computational costs. This paper explores a promising middle ground, where we employ a weak LLM that is significantly less resource-intensive than top-tier models, yet offers more automation than purely human feedback. We present a systematic study to evaluate and understand weak LLM's ability to generate feedback for alignment. Our empirical findings demonstrate that weak LLMs can provide feedback that rivals or even exceeds that of fully human-annotated data. Our study indicates a minimized impact of model size on feedback efficacy, shedding light on a scalable and sustainable alignment strategy. To deepen our understanding of alignment under weak LLM feedback, we conduct a series of qualitative and quantitative analyses, offering novel insights into the quality discrepancies between human feedback vs. weak LLM feedback. Code is publicly available at https://github.com/deeplearning-wisc/weakllmteacher.

3136. DaWin: Training-free Dynamic Weight Interpolation for Robust Adaptation

链接：https://iclr.cc/virtual/2025/poster/30016 abstract： Adapting a pre-trained foundation model on downstream tasks should ensure robustness against distribution shifts without the need to retrain the whole model. Although existing weight interpolation methods are simple yet effective, we argue their static nature limits downstream performance while achieving efficiency. In this work, we propose DaWin, a training-free dynamic weight interpolation method that leverages the entropy of individual models over each unlabeled test sample to assess model expertise, and compute per-sample interpolation coefficients dynamically. Unlike previous works that typically rely on additional training to learn such coefficients, our approach requires no training. Then, we propose a mixture modeling approach that greatly reduces inference overhead raised by dynamic interpolation. We validate DaWin on the large-scale visual recognition benchmarks, spanning 14 tasks across robust fine-tuning -- ImageNet and derived five distribution shift benchmarks -- and multi-task learning with eight classification tasks. Results demonstrate that DaWin achieves significant performance gain in considered settings, with minimal computational overhead. We further discuss DaWin's analytic behavior to explain its empirical success.

3137. CONDA: Adaptive Concept Bottleneck for Foundation Models Under Distribution Shifts

链接：https://iclr.cc/virtual/2025/poster/30736 abstract： Advancements in foundation models (FMs) have led to a paradigm shift in machinelearning. The rich, expressive feature representations from these pre-trained, large-scale FMs are leveraged for multiple downstream tasks, usually via lightweightfine-tuning of a shallow fully-connected network following the representation.However, the non-interpretable, black-box nature of this prediction pipeline can bea challenge, especially in critical domains, such as healthcare, finance, and security.In this paper, we explore the potential of Concept Bottleneck Models (CBMs)for transforming complex, non-interpretable foundation models into interpretabledecision-making pipelines using high-level concept vectors. Specifically, we focuson the test-time deployment of such an interpretable CBM pipeline “in the wild”,where the distribution of inputs often shifts from the original training distribution.We first identify the potential failure modes of such pipelines under different typesof distribution shifts. Then we propose an adaptive concept bottleneck frameworkto address these failure modes, that dynamically adapts the concept-vector bankand the prediction layer based solely on unlabeled

data from the target domain,without access to the source dataset. Empirical evaluations with various real-worlddistribution shifts show our framework produces concept-based interpretationsbetter aligned with the test data and boosts post-deployment accuracy by up to28%, aligning CBM performance with that of non-interpretable classification.

3138. Bi-Factorial Preference Optimization: Balancing Safety-Helpfulness in Language Models

链接：https://iclr.cc/virtual/2025/poster/30263 abstract： Fine-tuning large language models (LLMs) on human preferences, typically through reinforcement learning from human feedback (RLHF), has proven successful in enhancing their capabilities. However, ensuring the safety of LLMs during fine-tuning remains a critical concern, and mitigating the potential conflicts in safety and helpfulness is costly in RLHF. To address this issue, we propose a supervised learning framework called Bi-Factorial Preference Optimization (BFPO), which re-parameterizes a joint RLHF objective of both safety and helpfulness into a single supervised learning objective. In the supervised optimization, a labeling function is used to capture global preferences ranking to balance both safety and helpfulness. To evaluate BFPO, we develop a benchmark including comprehensive discriminative and generative tasks for helpfulness and harmlessness. The results indicate that our method significantly outperforms existing approaches in both safety and helpfulness. Moreover, BFPO eliminates the need for human prompting and annotation in LLM fine-tuning while achieving the same level of safety as methods that heavily rely on human labor, with less than 10\% of the computational resources. The training recipes and models will be released.

3139. LongGenBench: Benchmarking Long-Form Generation in Long Context LLMs

链接：https://iclr.cc/virtual/2025/poster/31092 abstract： Current benchmarks like ``$\textit{Needle-in-a-Haystack}$" ($\textit{NIAH}$), $\textit{Ruler}$, and $\textit{Needlebench}$ focus on models' ability to understand long-context input sequences but fail to capture a critical dimension: the generation of high-quality long-form text. Applications such as design proposals, technical documentation, and creative writing rely on coherent, instruction-following outputs over extended sequences—a challenge that existing benchmarks do not adequately address. To fill this gap, we introduce $\textit{LongGenBench}$, a novel benchmark designed to rigorously evaluate large language models' (LLMs) ability to generate long text while adhering to complex instructions. Through tasks requiring specific events or constraints within generated text, $\textit{LongGenBench}$ evaluates model performance across four distinct scenarios, three instruction types, and two generation-lengths (16K and 32K tokens). Our evaluation of ten state-of-the-art LLMs reveals that, despite strong results on $\textit{Ruler}$, all models struggled with long text generation on $\textit{LongGenBench}$, particularly as text length increased. This suggests that current LLMs are not yet equipped to meet the demands of real-world, long-form text generation. We open-source $\textit{LongGenBench}$ to promote comprehensive evaluation and improvement in this critical area, with code and data available at ${anonymousurl}$.

3140. Learning Graph Invariance by Harnessing Spuriosity

链接：https://iclr.cc/virtual/2025/poster/29448 abstract： Recently, graph invariant learning has become the *de facto* approach to tackle the Out-of-Distribution (OOD) generalization failure in graph representation learning. They generically follow the framework of invariant risk minimization to capture the invariance of graph data from different environments. Despite some success, it remains unclear to what extent existing approaches have captured invariant features for OOD generalization on graphs. In this work, we find that representative OOD methods such as IRM and VRex, and their variants on graph invariant learning may have captured a limited set of invariant features. To tackle this challenge, we propose $\texttt{LIRS}$, a novel learning framework designed to **L**earn graph **I**nvariance by **R**emoving **S**purious features. Different from most existing approaches that *directly* learn the invariant features, $\texttt{LIRS}$ takes an *indirect* approach by first learning the spurious features and then removing them from the ERM-learned features, which contains both spurious and invariant features. We demonstrate that learning the invariant graph features in an *indirect* way can learn a more comprehensive set of invariant features. Moreover, our proposed method outperforms the second-best method by as much as 25.50% across all competitive baseline methods, highlighting its effectiveness in learning graph invariant features.

3141. Zigzag Diffusion Sampling: Diffusion Models Can Self-Improve via Self-Reflection

链接：https://iclr.cc/virtual/2025/poster/29938 abstract： Diffusion models, the most popular generative paradigm so far, can inject conditional information into the generation path to guide the latent towards desired directions. However, existing text-to-image diffusion models often fail to maintain high image quality and high prompt-image alignment for those challenging prompts. To mitigate this issue and enhance existing pretrained diffusion models, we mainly made three contributions in this paper. First, we propose diffusion self-reflection that alternately performs denoising and inversion and demonstrate that such diffusion self-reflection can leverage the guidance gap between denoising and inversion to capture prompt-related semantic information with theoretical and empirical evidence. Second, motivated by theoretical analysis, we derive Zigzag Diffusion Sampling (Z-Sampling), a novel self-reflection-based diffusion sampling method that leverages the guidance gap between denosing and inversion to accumulate semantic information step by step along the sampling path, leading to improved sampling results. Moreover, as a plug-and-play method, Z-Sampling can be generally applied to various diffusion models (e.g., accelerated ones and Transformer-based ones) with very limited coding and computational costs. Third, our extensive experiments demonstrate

that Z-Sampling can generally and significantly enhance generation quality across various benchmark datasets, diffusion models, and performance evaluation metrics. For example, DreamShaper with Z-Sampling can self-improve with the HPSv2 winning rate up to 94% over the original results. Moreover, Z-Sampling can further enhance existing diffusion models combined with other orthogonal methods, including Diffusion-DPO. The code is publicly available atgithub.com/xie-lab-ml/Zigzag-Diffusion-Sampling.

3142. Grounding Continuous Representations in Geometry: Equivariant Neural Fields

链接：https://iclr.cc/virtual/2025/poster/30651 abstract： Conditional Neural Fields (CNFs) are increasingly being leveraged as continuous signal representations, by associating each data-sample with a latent variable that conditions a shared backbone Neural Field (NeF) to reconstruct the sample. However, existing CNF architectures face limitations when using this latent downstream in tasks requiring fine-grained geometric reasoning, such as classification and segmentation. We posit that this results from lack of explicit modelling of geometric information (e.g. locality in the signal or the orientation of a feature) in the latent space of CNFs. As such, we propose Equivariant Neural Fields (ENFs), a novel CNF architecture which uses a geometry-informed cross-attention to condition the NeF on a geometric variable—a latent point cloud of features—that enables an equivariant decoding from latent to field. We show that this approach induces a steerability property by which both field and latent are grounded in geometry and amenable to transformation laws: if the field transforms, the latent representation transforms accordingly—and vice versa. Crucially, this equivariance relation ensures that the latent is capable of (1) representing geometric patterns faitfhully, allowing for geometric reasoning in latent space, (2) weight-sharing over similar local patterns, allowing for efficient learning of datasets of fields. We validate these main properties in a range of tasks including classification, segmentation, forecasting, reconstruction and generative modelling, showing clear improvement over baselines with a geometry-free latent space.

3143. Shedding Light on Time Series Classification using Interpretability Gated Networks

链接：https://iclr.cc/virtual/2025/poster/28432 abstract： In time-series classification, interpretable models can bring additional insights but be outperformed by deep models since human-understandable features have limited expressivity and flexibility. In this work, we present InterpGN, a framework that integrates an interpretable model and a deep neural network. Within this framework, we introduce a novel gating function design based on the confidence of the interpretable expert, preserving interpretability for samples where interpretable features are significant while also identifying samples that require additional expertise. For the interpretable expert, we incorporate shapelets to effectively model shape-level features for time-series data. We introduce a variant of Shapelet Transforms to build logical predicates using shapelets. Our proposed model achieves comparable performance with state-of-the-art deep learning models while additionally providing interpretable classifiers for various benchmark datasets. We further show that our models improve on quantitative shapelet quality and interpretability metrics over existing shapelet-learning formulations. Finally, we show that our models can integrate additional advanced architectures and be applied to real-world tasks beyond standard benchmarks such as the MIMIC-III and time series extrinsic regression datasets.

3144. Measuring And Improving Engagement of Text-to-Image Generation Models

链接：https://iclr.cc/virtual/2025/poster/32082 abstract： Recent advances in text-to-image generation have achieved impressive aesthetic quality, making these models usable for both personal and commercial purposes. However, in the fields of marketing and advertising, images are often created to be more engaging, as reflected in user behaviors such as increasing clicks, likes, and purchases, in addition to being aesthetically pleasing. To this end, we introduce the challenge of optimizing the image generation process for improved viewer engagement. In order to study image engagement and utility in real-world marketing scenarios, we collect EngagingImageNet, the first large-scale dataset of images, along with associated user engagement metrics. Further, we find that existing image evaluation metrics like aesthetics, CLIPScore, PickScore, ImageReward, etc. are unable to capture viewer engagement. To address the lack of reliable metrics for assessing image utility, we use the EngagingImageNet dataset to train EngageNet, an engagement-aware Vision Language Model (VLM) that predicts viewer engagement of images by leveraging contextual information about the tweet content, enterprise details, and posting time. We then explore methods to enhance the engagement of text-to-image models, making initial strides in this direction. These include conditioning image generation on improved prompts, supervised fine-tuning of stable diffusion on high-performing images, and reinforcement learning to align stable diffusion with EngageNet-based reward signals, all of which lead to the generation of images with higher viewer engagement. Finally, we propose the Engagement Arena, to benchmark text-to-image models based on their ability to generate engaging images, using EngageNet as the evaluator, thereby encouraging the research community to measure further advances in the engagement of text-to-image modeling. These contributions provide a new pathway for advancing utility-driven image generation, with significant implications for the commercial application of image generation. We have released our code and dataset on behavior-in-the-wild.github.io/image-engagement.

3145. Discriminating image representations with principal distortions

链接：https://iclr.cc/virtual/2025/poster/27965 abstract： Image representations (artificial or biological) are often compared in terms of their global geometric structure; however, representations with similar global structure can have strikingly different local geometries. Here, we propose a framework for comparing a set of image representations in terms of their local geometries. We quantify the local geometry of a representation using the Fisher information matrix, a standard statistical tool for characterizing the sensitivity to local stimulus distortions, and use this as a substrate for a metric on the local geometry in the vicinity of a base image. This metric may then be used to optimally differentiate a set of models, by finding a pair of "principal distortions" that maximize the variance of the models under this metric. As an example, we use this framework to compare a set of simple models of the early visual system, identifying a novel set of image distortions that allow immediate comparison of the models by visual inspection. In a second example, we apply our method to a set of deep neural network models and reveal differences in the local geometry that arise due to architecture and training types. These examples demonstrate how our framework can be used to probe for informative differences in local sensitivities between complex models, and suggest how it could be used to compare model representations with human perception.

3146. MetaDesigner: Advancing Artistic Typography through AI-Driven, User-Centric, and Multilingual WordArt Synthesis

链接：https://iclr.cc/virtual/2025/poster/29912 abstract： MetaDesigner introduces a transformative framework for artistic typography synthesis, powered by Large Language Models (LLMs) and grounded in a user-centric design paradigm. Its foundation is a multi-agent system comprising the Pipeline, Glyph, and Texture agents, which collectively orchestrate the creation of customizable WordArt, ranging from semantic enhancements to intricate textural elements. A central feedback mechanism leverages insights from both multimodal models and user evaluations, enabling iterative refinement of design parameters. Through this iterative process, MetaDesigner dynamically adjusts hyperparameters to align with user-defined stylistic and thematic preferences, consistently delivering WordArt that excels in visual quality and contextual resonance. Empirical evaluations underscore the system's versatility and effectiveness across diverse WordArt applications, yielding outputs that are both aesthetically compelling and context-sensitive.

3147. One-Prompt-One-Story: Free-Lunch Consistent Text-to-Image Generation Using a Single Prompt

链接：https://iclr.cc/virtual/2025/poster/29063 abstract：

3148. LLM-SR: Scientific Equation Discovery via Programming with Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28487 abstract：

3149. Advancing Prompt-Based Methods for Replay-Independent General Continual Learning

链接：https://iclr.cc/virtual/2025/poster/29430 abstract： General continual learning (GCL) is a broad concept to describe real-world continual learning (CL) problems, which are often characterized by online data streams without distinct transitions between tasks, i.e., blurry task boundaries. Such requirements result in poor initial performance, limited generalizability, and severe catastrophic forgetting, heavily impacting the effectiveness of mainstream GCL models trained from scratch. While the use of a frozen pretrained backbone with appropriate prompt tuning can partially address these challenges, such prompt-based methods remain suboptimal for CL of remaining tunable parameters on the fly. In this regard, we propose an innovative approach named MISA (Mask and Initial Session Adaption) to advance prompt-based methods in GCL. It includes a forgetting-aware initial session adaption that employs pretraining data to initialize prompt parameters and improve generalizability, as well as a non-parametric logit mask of the output layers to mitigate catastrophic forgetting. Empirical results demonstrate substantial performance gains of our approach compared to recent competitors, especially without a replay buffer (e.g., up to 18.39, 22.06, and 11.96% points performance lead on CIFAR-100, Tiny-ImageNet, and ImageNet-R, respectively). Moreover, our approach features the plug-in nature for prompt-based methods, independence of replay, ease of implementation, and avoidance of CL-relevant hyperparameters, serving as a strong baseline for GCL research. Our source code is publicly available at https://github.com/kangzhiq/MISA

3150. FACTS: A Factored State-Space Framework for World Modelling

链接：https://iclr.cc/virtual/2025/poster/28966 abstract： World modelling is essential for understanding and predicting the dynamics of complex systems by learning both spatial and temporal dependencies. However, current frameworks, such as Transformers and selective state-space models like Mambas, exhibit limitations in efficiently encoding spatial and temporal structures, particularly in scenarios requiring long-term high-dimensional sequence modelling. To address these issues, we propose a novel recurrent framework, the FACTored State-space (FACTS) model, for spatial-temporal world modelling. The FACTS framework constructs a graph-structured memory with a routing mechanism that learns permutable memory representations, ensuring invariance to input permutations while adapting through selective state-space propagation.

Furthermore, FACTS supports parallel computation of high-dimensional sequences. We empirically evaluate FACTS across diverse tasks, including multivariate time series forecasting, object-centric world modelling, and spatial-temporal graph prediction, demonstrating that it consistently outperforms or matches specialised state-of-the-art models, despite its general-purpose world modelling design.

3151. Wasserstein Distances, Neuronal Entanglement, and Sparsity

链接：https://iclr.cc/virtual/2025/poster/29028 abstract： Disentangling polysemantic neurons is at the core of many current approaches to interpretability of large language models. Here we attempt to study how disentanglement can be used to understand performance, particularly under weight sparsity, a leading post-training optimization technique. We suggest a novel measure for estimating neuronal entanglement: the Wasserstein distance of a neuron's output distribution to a Gaussian. Moreover, we show the existence of a small number of highly entangled "Wasserstein Neurons" in each linear layer of an LLM, characterized by their highly non-Gaussian output distributions, their role in mapping similar inputs to dissimilar outputs, and their significant impact on model accuracy. To study these phenomena, we propose a new experimental framework for disentangling polysemantic neurons. Our framework separates each layer's inputs to create a mixture of experts where each neuron's output is computed by a mixture of neurons of lower Wasserstein distance, each better at maintaining accuracy when sparsified without retraining. We provide strong evidence that this is because the mixture of sparse experts is effectively disentangling the input-output relationship of individual neurons, in particular the difficult Wasserstein neurons.

3152. Lean-STaR: Learning to Interleave Thinking and Proving

链接：https://iclr.cc/virtual/2025/poster/29611 abstract： Traditional language model-based theorem proving assumes that by training on a sufficient amount of formal proof data, a model will learn to prove theorems. Our key observation is that a wealth of informal information that is not present in formal proofs can be useful for learning to prove theorems. For instance, humans think through steps of a proof, but this thought process is not visible in the resulting code. We present Lean-STaR, a framework for training language models to produce informal thoughts prior to each step of a proof, thereby boosting the model's theorem-proving capabilities. Lean-STaR uses retrospective ground-truth tactics to generate synthetic thoughts for training the language model. At inference time, the trained model directly generates the thoughts prior to the prediction of the tactics in each proof step. Building on the self-taught reasoner framework, we then apply expert iteration to further fine-tune the model on the correct proofs it samples and verifies using the Lean solver. Lean-STaR significantly outperform base models (43.4% → 46.3%, Pass@64). We also analyze the impact of the augmented thoughts on various aspects of the theorem proving process, providing insights into their effectiveness.

3153. Self-Play Preference Optimization for Language Model Alignment

链接：https://iclr.cc/virtual/2025/poster/29189 abstract： Standard reinforcement learning from human feedback (RLHF) approaches relying on parametric models like the Bradley-Terry model fall short in capturing the intransitivity and irrationality in human preferences. Recent advancements suggest that directly working with preference probabilities can yield a more accurate reflection of human preferences, enabling more flexible and accurate language model alignment. In this paper, we propose a self-play-based method for language model alignment, which treats the problem as a constant-sum two-player game aimed at identifying the Nash equilibrium policy. Our approach, dubbed Self-Play Preference Optimization (SPPO), utilizes iterative policy updates to provably approximate the Nash equilibrium. Additionally, we propose a new SPPO objective which is both strongly motivated by theory and is simple and effective in practice.In our experiments, using only 60k prompts (without responses) from the UltraFeedback dataset and without any prompt augmentation, by leveraging a pre-trained preference model PairRM with only 0.4B parameters, SPPO can obtain a model from fine-tuning Mistral-7B-Instruct-v0.2 that achieves the state-of-the-art length-controlled win-rate of 28.53\% against GPT-4-Turbo on AlpacaEval 2.0. It also outperforms the (iterative) DPO and IPO on MT-Bench, Arena-Hard, and the Open LLM Leaderboard.Starting from a stronger base model Llama-3-8B-Instruct, we are able to achieve a length-controlled win rate of 38.77\%.Notably, the strong performance of SPPO is achieved without additional external supervision (e.g., responses, preferences, etc.) from GPT-4 or other stronger language models.

3154. Inference Scaling Laws: An Empirical Analysis of Compute-Optimal Inference for LLM Problem-Solving

链接：https://iclr.cc/virtual/2025/poster/29417 abstract： While the scaling laws of large language models (LLMs) training have been extensively studied, optimal inference configurations of LLMs remain underexplored. We study inference scaling laws (aka test-time scaling laws) and compute-optimal inference, focusing on the trade-offs between model sizes and generating additional tokens with different inference strategies. As a first step towards understanding and designing compute-optimal inference methods, we studied cost-performance trade-offs for inference strategies such as greedy search, majority voting, best-of-n, weighted voting, and two different tree search algorithms, using different model sizes and compute budgets. Our findings suggest that scaling inference compute with inference strategies can be more computationally efficient than scaling model parameters. Additionally, smaller models combined with advanced inference algorithms offer Pareto-optimal trade-offs in cost and performance. For example, the Llemma-7B model, when paired with our novel tree search algorithm, consistently outperforms the Llemma-34B model across all tested inference strategies on the MATH benchmark. We hope these insights contribute to a deeper understanding of inference scaling laws (test-time scaling laws) for LLMs.

3155. OpenRCA: Can Large Language Models Locate the Root Cause of Software Failures?

链接：https://iclr.cc/virtual/2025/poster/32093 abstract： Large language models (LLMs) are driving substantial advancements in software engineering, with successful applications like Copilot and Cursor transforming real-world development practices. However, current research predominantly focuses on the early stages of development, such as code generation, while overlooking the post-development phases that are crucial to user experience. To explore the potential of LLMs in this direction, we propose OpenRCA, a benchmark dataset and evaluation framework for assessing LLMs' ability to identify the root cause of software failures. OpenRCA includes 335 failures from three enterprise software systems, along with over 68 GB of telemetry data (logs, metrics, and traces). Given a failure case and its associated telemetry, the LLM is tasked to identify the root cause that triggered the failure, requiring comprehension of software dependencies and reasoning over heterogeneous, long-context telemetry data. Our results show substantial room for improvement, as current models can only handle the simplest cases. Even with the specially designed RCA-agent, the best-performing model, Claude 3.5, solved only 11.34% failure cases. Our work paves the way for future research in this direction.

3156. Flow Matching with General Discrete Paths: A Kinetic-Optimal Perspective

链接：https://iclr.cc/virtual/2025/poster/28039 abstract： The design space of discrete-space diffusion or flow generative models are significantly less well-understood than their continuous-space counterparts, with many works focusing only on a simple masked construction.In this work, we aim to take a holistic approach to the construction of discrete generative models based on continuous-time Markov chains, and for the first time, allow the use of arbitrary discrete probability paths, or colloquially, corruption processes. Through the lens of optimizing the symmetric kinetic energy, we propose velocity formulas that can be applied to any given probability path, completely decoupling the probability and velocity, and giving the user the freedom to specify any desirable probability path based on expert knowledge specific to the data domain. Furthermore, we find that a special construction of mixture probability paths optimizes the symmetric kinetic energy for the discrete case.We empirically validate the usefulness of this new design space across multiple modalities: text generation, inorganic material generation, and image generation. We find that we can outperform the mask construction even in text with kinetic-optimal mixture paths, while we can make use of domain-specific constructions of the probability path over the visual domain.

3157. MP-Mat: A 3D-and-Instance-Aware Human Matting and Editing Framework with Multiplane Representation

链接：https://iclr.cc/virtual/2025/poster/29283 abstract： Human instance matting aims to estimate an alpha matte for each human instance in an image, which is challenging as it easily fails in complex cases requiring disentangling mingled pixels belonging to multiple instances along hairy and thin boundary structures. In this work, we address this by introducing MP-Mat, a novel 3D-and-instance-aware matting framework with multiplane representation, where the multiplane concept is designed from two different perspectives: scene geometry level and instance level. Specifically, we first build feature-level multiplane representations to split the scene into multiple planes based on depth differences. This approach makes the scene representation 3D-aware, and can serve as an effective clue for splitting instances in different 3D positions, thereby improving interpretability and boundary handling ability especially in occlusion areas. Then, we introduce another multiplane representation that splits the scene in an instance-level perspective, and represents each instance with both matte and color. We also treat background as a special instance, which is often overlooked by existing methods. Such an instance-level representation facilitates both foreground and background content awareness, and is useful for other down-stream tasks like image editing. Once built, the representation can be reused to realize controllable instance-level image editing with high efficiency. Extensive experiments validate the clear advantage of MP-Mat in matting task. We also demonstrate its superiority in image editing tasks, an area under-explored by existing matting-focused methods, where our approach under zero-shot inference even outperforms trained specialized image editing techniques by large margins. Code is open-sourced at https://github.com/JiaoSiyi/MPMat.git.

3158. Enhancing Zeroth-order Fine-tuning for Language Models with Low-rank Structures

链接：https://iclr.cc/virtual/2025/poster/30716 abstract： Parameter-efficient fine-tuning (PEFT) significantly reduces memory costs when adapting large language models (LLMs) for downstream applications. However, traditional first-order (FO) fine-tuning algorithms incur substantial memory overhead due to the need to store activation values for back-propagation during gradient computation, particularly in long-context fine-tuning tasks. Zeroth-order (ZO) algorithms offer a promising alternative by approximating gradients using finite differences of function values, thus eliminating the need for activation storage. Nevertheless, existing ZO methods struggle to capture the low-rank gradient structure common in LLM fine-tuning, leading to suboptimal performance. This paper proposes a low-rank ZO gradient estimator and introduces a novel low-rank ZO algorithm (LOZO) that effectively captures this structure in LLMs. We provide convergence guarantees for LOZO by framing it as a subspace optimization method. Additionally, its low-rank nature enables LOZO to integrate with momentum techniques while incurring negligible extra memory costs. Extensive experiments across various model sizes and downstream tasks demonstrate that LOZO and its momentum-based variant outperform existing ZO methods and closely approach the performance of FO algorithms.

3159. On the Performance Analysis of Momentum Method: A Frequency Domain Perspective

链接：https://iclr.cc/virtual/2025/poster/28003 abstract： Momentum-based optimizers are widely adopted for training neural networks. However, the optimal selection of momentum coefficients remains elusive. This uncertainty impedes a clear understanding of the role of momentum in stochastic gradient methods. In this paper, we present a frequency domain analysis framework that interprets the momentum method as a time-variant filter for gradients, where adjustments to momentum coefficients modify the filter characteristics. Our experiments support this perspective and provide a deeper understanding of the mechanism involved. Moreover, our analysis reveals the following significant findings: high-frequency gradient components are undesired in the late stages of training; preserving the original gradient in the early stages, and gradually amplifying low-frequency gradient components during training both enhance performance. Based on these insights, we propose Frequency Stochastic Gradient Descent with Momentum (FSGDM), a heuristic optimizer that dynamically adjusts the momentum filtering characteristic with an empirically effective dynamic magnitude response. Experimental results demonstrate the superiority of FSGDM over conventional momentum optimizers.

3160. Brain Mapping with Dense Features: Grounding Cortical Semantic Selectivity in Natural Images With Vision Transformers

链接：https://iclr.cc/virtual/2025/poster/27735 abstract： We introduce BrainSAIL (Semantic Attribution and Image Localization), a method for linking neural selectivity with spatially distributed semantic visual concepts in natural scenes. BrainSAIL leverages recent advances in large-scale artificial neural networks, using them to provide insights into the functional topology of the brain. To overcome the challenge presented by the co-occurrence of multiple categories in natural images, BrainSAIL exploits semantically consistent, dense spatial features from pre-trained vision models, building upon their demonstrated ability to robustly predict neural activity. This method derives clean, spatially dense embeddings without requiring any additional training, and employs a novel denoising process that leverages the semantic consistency of images under random augmentations. By unifying the space of whole-image embeddings and dense visual features and then applying voxel-wise encoding models to these features, we enable the identification of specific subregions of each image which drive selectivity patterns in different areas of the higher visual cortex. This provides a powerful tool for dissecting the neural mechanisms that underlie semantic visual processing for natural images. We validate BrainSAIL on cortical regions with known category selectivity, demonstrating its ability to accurately localize and disentangle selectivity to diverse visual concepts. Next, we demonstrate BrainSAIL's ability to characterize high-level visual selectivity to scene properties and low-level visual features such as depth, luminance, and saturation, providing insights into the encoding of complex visual information. Finally, we use BrainSAIL to directly compare the feature selectivity of different brain encoding models across different regions of interest in visual cortex. Our innovative method paves the way for significant advances in mapping and decomposing high-level visual representations in the human brain.

3161. SaMer: A Scenario-aware Multi-dimensional Evaluator for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29184 abstract： Evaluating the response quality of large language models (LLMs) for open-ended questions poses a significant challenge, especially given the subjectivity and multi-dimensionality of "quality" in natural language generation. Existing LLM evaluators often neglect that different scenarios require distinct evaluation criteria. In this work, we propose SaMer, a scenario-aware multi-dimensional evaluator designed to provide both overall and fine-grained assessments of LLM-generated responses. Unlike fixed-dimension evaluation approaches, SaMer adapts to different scenarios by automatically identifying and prioritizing relevant evaluation dimensions tailored to the given query. To achieve this, we construct a large-scale fine-grained preference dataset spanning multiple real-world scenarios, each with distinct evaluation dimensions. We then leverage a text embedding model combined with three specialized heads to predict the appropriate evaluation dimensions and corresponding scores, as well as the respective weights that contribute to the overall score. The resulting model offers fine-grained and interpretable evaluations and shows robust adaptability across diverse scenarios. Extensive experiments on eight single rating and pairwise comparison datasets demonstrate that SaMer outperforms existing baselines in a variety of evaluation tasks, showcasing its robustness, versatility, and generalizability.

3162. Self-supervised Monocular Depth Estimation Robust to Reflective Surface Leveraged by Triplet Mining

链接：https://iclr.cc/virtual/2025/poster/29298 abstract： Self-supervised monocular depth estimation (SSMDE) aims to predict the dense depth map of a monocular image, by learning depth from RGB image sequences, eliminating the need for ground-truth depth labels.Although this approach simplifies data acquisition compared to supervised methods, it struggles with reflective surfaces, as they violate the assumptions of Lambertian reflectance, leading to inaccurate training on such surfaces.To tackle this problem, we propose a novel training strategy for an SSMDE by leveraging triplet mining to pinpoint reflective regions at the pixel level, guided by the camera geometry between different viewpoints.The proposed reflection-aware triplet mining loss specifically penalizes the inappropriate photometric error minimization on the localized reflective regions while preserving depth accuracy on non-reflective areas.We also incorporate a reflection-aware knowledge distillation method that enables a student model to selectively learn the pixel-level knowledge from reflective and non-reflective regions. This results in robust depth

estimation across areas.Evaluation results on multiple datasets demonstrate that our method effectively enhances depth quality on reflective surfaces and outperforms state-of-the-art SSMDE baselines.

3163. InstantSplamp: Fast and Generalizable Stenography Framework for Generative Gaussian Splatting

链接：https://iclr.cc/virtual/2025/poster/27757 abstract： With the rapid development of large generative models for 3D, especially the evolution from NeRF representations to more efficient Gaussian Splatting, the synthesis of 3D assets has become increasingly fast and efficient, enabling the large-scale publication and sharing of generated 3D objects. However, while existing methods can add watermarks or steganographic information to individual 3D assets, they often require time-consuming per-scene training and optimization, leading to watermarking overheads that can far exceed the time required for asset generation itself, making deployment impractical for generating large collections of 3D objects. To address this, we propose InstantSplamp a framework that seamlessly integrates the 3D steganography pipeline into large 3D generative models without introducing explicit additional time costs. Guided by visual foundation models,InstantSplamp subtly injects hidden information like copyright tags during asset generation, enabling effective embedding and recovery of watermarks within generated 3D assets while preserving original visual quality. Experiments across various potential deployment scenarios demonstrate that \model~strikes an optimal balance between rendering quality and hiding fidelity, as well as between hiding performance and speed. Compared to existing per-scene optimization techniques for 3D assets, InstantSplamp reduces their watermarking training overheads that are multiples of generation time to nearly zero, paving the way for real-world deployment at scale. Project page: https://gaussian-stego.github.io/.

3164. PFDiff: Training-Free Acceleration of Diffusion Models Combining Past and Future Scores

链接：https://iclr.cc/virtual/2025/poster/27822 abstract： Diffusion Probabilistic Models (DPMs) have shown remarkable potential in image generation, but their sampling efficiency is hindered by the need for numerous denoising steps. Most existing solutions accelerate the sampling process by proposing fast ODE solvers. However, the inevitable discretization errors of the ODE solvers are significantly magnified when the number of function evaluations (NFE) is fewer. In this work, we propose PFDiff, a novel training-free and orthogonal timestep-skipping strategy, which enables existing fast ODE solvers to operate with fewer NFE. Specifically, PFDiff initially utilizes score replacement from past time steps to predict a springboard. Subsequently, it employs this ``springboard'' along with foresight updates inspired by Nesterov momentum to rapidly update current intermediate states. This approach effectively reduces unnecessary NFE while correcting for discretization errors inherent in first-order ODE solvers. Experimental results demonstrate that PFDiff exhibits flexible applicability across various pre-trained DPMs, particularly excelling in conditional DPMs and surpassing previous state-of-the-art training-free methods. For instance, using DDIM as a baseline, we achieved 16.46 FID (4 NFE) compared to 138.81 FID with DDIM on ImageNet 64x64 with classifier guidance, and 13.06 FID (10 NFE) on Stable Diffusion with 7.5 guidance scale. Code is available at https://github.com/onefly123/PFDiff.

3165. 4K4DGen: Panoramic 4D Generation at 4K Resolution

链接：https://iclr.cc/virtual/2025/poster/28230 abstract： The blooming of virtual reality and augmented reality (VR/AR) technologies has driven an increasing demand for the creation of high-quality, immersive, and dynamic environments. However, existing generative techniques either focus solely on dynamic objects or perform outpainting from a single perspective image, failing to meet the requirements of VR/AR applications that need free-viewpoint, 360° virtual views where users can move in all directions. In this work, we tackle the challenging task of elevating a single panorama to an immersive 4D experience. For the first time, we demonstrate the capability to generate omnidirectional dynamic scenes with 360° views at 4K (4096 $\times$ 2048) resolution, thereby providing an immersive user experience. Our method introduces a pipeline that facilitates natural scene animations and optimizes a set of 3D Gaussians using efficient splatting techniques for real-time exploration. To overcome the lack of scene-scale annotated 4D data and models, especially in panoramic formats, we propose a novel Panoramic Denoiser that adapts generic 2D diffusion priors to animate consistently in 360° images, transforming them into panoramic videos with dynamic scenes at targeted regions. Subsequently, we propose Dynamic Panoramic Lifting to elevate the panoramic video into a 4D immersive environment while preserving spatial and temporal consistency. By transferring prior knowledge from 2D models in the perspective domain to the panoramic domain and the 4D lifting with spatial appearance and geometry regularization, we achieve high-quality Panorama-to-4D generation at a resolution of 4K for the first time. Project page: https://4k4dgen.github.io/.

3166. Fast and Slow Streams for Online Time Series Forecasting Without Information Leakage

链接：https://iclr.cc/virtual/2025/poster/30196 abstract： Current research in online time series forecasting (OTSF) faces two significant issues. The first is information leakage, where models make predictions and are then evaluated on historical time steps that have already been used in backpropagation for parameter updates. The second is practicality: while forecasting in real-world applications typically emphasizes looking ahead and anticipating future uncertainties, prediction sequences in this setting include only one future step with the remaining being observed time points. This necessitates a redefinition of the OTSF setting, focusing on predicting unknown future steps and evaluating unobserved data points. Following this new setting,

challenges arise in leveraging incomplete pairs of ground truth and predictions for backpropagation, as well as in generalizing accurate information without overfitting to noise from recent data streams. To address these challenges, we propose a novel dual-stream framework for online forecasting (DSOF): a slow stream that updates with complete data using experience replay, and a fast stream that adapts to recent data through temporal difference learning. This dual-stream approach updates a teacher-student model learned through a residual learning strategy, generating predictions in a coarse-to-fine manner. Extensive experiments demonstrate its improvement in forecasting performance in changing environments. Our code is publicly available at https://github.com/yyalau/iclr2025_dsof.

3167. FLIP: Flow-Centric Generative Planning as General-Purpose Manipulation World Model

链接：https://iclr.cc/virtual/2025/poster/30597 abstract： We aim to develop a model-based planning framework for world models that can be scaled with increasing model and data budgets for general-purpose manipulation tasks with only language and vision inputs. To this end, we present FLow-Centrlc generative Planning (FLIP), a model-based planning algorithm on visual space that features three key modules: 1) a multi-modal flow generation model as the general-purpose action proposal module; 2) a flow-conditioned video generation model as the dynamics module; and 3) a vision-language representation learning model as the value module. Given an initial image and language instruction as the goal, FLIP can progressively search for long-horizon flow and video plans that maximize the discounted return to accomplish the task. FLIP is able to synthesize long-horizon plans across objects, robots, and tasks with image flows as the general action representation, and the dense flow information also provides rich guidance for long-horizon video generation. In addition, the synthesized flow and video plans can guide the training of low-level control policies for robot execution. Experiments on diverse benchmarks demonstrate that FLIP can improve both the success rates and quality of long-horizon video plan synthesis and has the interactive world model property, opening up wider applications for future works. Video demos are on our website: https://nus-lins-lab.github.io/flipweb/.

3168. SciImage: How good are multimodal large language models at scientific text-to-image generation?

链接：https://iclr.cc/virtual/2025/poster/27964 abstract： Multimodal large language models (LLMs) have demonstrated impressive capabilities in generating high-quality images from textual instructions. However, their performance in generating scientific images—a critical application for accelerating scientific progress—remains underexplored. In this work, we address this gap by introducing SciImage, a benchmark designed to evaluate the multimodal capabilities of LLMs in generating scientific images from textual descriptions. SciImage assesses three key dimensions of understanding: spatial, numeric, and attribute comprehension, as well as their combinations, focusing on the relationships between scientific objects (e.g., squares, circles). We evaluate seven models, GPT-4o, Llama, AutomaTikZ, Dall-E, StableDiffusion, GPT-o1 and Qwen2.5-Coder-Instruct using two modes of output generation: code-based outputs (Python, TikZ) and direct raster image generation. Additionally, we examine four different input languages: English, German, Farsi, and Chinese. Our evaluation, conducted with 11 scientists across three criteria (correctness, relevance, and scientific accuracy), reveals that while GPT4-o produces outputs of decent quality for simpler prompts involving individual dimensions such as spatial, numeric, or attribute understanding in isolation, all models face challenges in this task, especially for more complex prompts. SciImage is available: huggingface.co/datasets/casszhao/SciImage

3169. Binary Losses for Density Ratio Estimation

链接：https://iclr.cc/virtual/2025/poster/30970 abstract： Estimating the ratio of two probability densities from a finite number of observations is a central machine learning problem. A common approach is to construct estimators using binary classifiers that distinguish observations from the two densities. However, the accuracy of these estimators depends on the choice of the binary loss function, raising the question of which loss function to choose based on desired error properties. For example, traditional loss functions, such as logistic or boosting loss, prioritize accurate estimation of small density ratio values over large ones, even though the latter are more critical in many applications.In this work, we start with prescribed error measures in a class of Bregman divergences and characterize all loss functions that result in density ratio estimators with small error. Our characterization extends results on composite binary losses from Reid & Williamson (2010) and their connection to density ratio estimation as identified by Menon & Ong (2016). As a result, we obtain a simple recipe for constructing loss functions with certain properties, such as those that prioritize an accurate estimation of large density ratio values. Our novel loss functions outperform related approaches for resolving parameter choice issues of 11 deep domain adaptation algorithms in average performance across 484 real-world tasks including sensor signals, texts, and images.

3170. Data Taggants: Dataset Ownership Verification Via Harmless Targeted Data Poisoning

链接：https://iclr.cc/virtual/2025/poster/30866 abstract：

3171. DICE: End-to-end Deformation Capture of Hand-Face Interactions from a Single Image

链接：https://iclr.cc/virtual/2025/poster/28179 abstract：

3172. BANGS: Game-theoretic Node Selection for Graph Self-Training

链接：https://iclr.cc/virtual/2025/poster/28782 abstract： Graph self-training is a semi-supervised learning method that iteratively selects a set of unlabeled data to retrain the underlying graph neural network (GNN) model and improve its prediction performance. While selecting highly confident nodes has proven effective for self-training, this pseudo-labeling strategy ignores the combinatorial dependencies between nodes and suffers from a local view of the distribution.To overcome these issues, we propose BANGS, a novel framework that unifies the labeling strategy with conditional mutual information as the objective of node selection. Our approach---grounded in game theory---selects nodes in a combinatorial fashion and provides theoretical guarantees for robustness under noisy objective. More specifically, unlike traditional methods that rank and select nodes independently, BANGS considers nodes as a collective set in the self-training process. Our method demonstrates superior performance and robustness across various datasets, base models, and hyperparameter settings, outperforming existing techniques. The codebase is available on https://anonymous.4open.science/r/BANGS-3EA4.

3173. CityAnchor: City-scale 3D Visual Grounding with Multi-modality LLMs

链接：https://iclr.cc/virtual/2025/poster/30802 abstract： In this paper, we present a 3D visual grounding method called CityAnchor for localizing an urban object in a city-scale point cloud. Recent developments in multiview reconstruction enable us to reconstruct city-scale point clouds but how to conduct visual grounding on such a large-scale urban point cloud remains an open problem. Previous 3D visual grounding system mainly concentrates on localizing an object in an image or a small-scale point cloud, which is not accurate and efficient enough to scale up to a city-scale point cloud. We address this problem with a multi-modality LLM which consists of two stages, a coarse localization and a fine-grained matching. Given the text descriptions, the coarse localization stage locates possible regions on a projected 2D map of the point cloud while the fine-grained matching stage accurately determines the most matched object in these possible regions. We conduct experiments on the CityRefer dataset and a new synthetic dataset annotated by us, both of which demonstrate our method can produce accurate 3D visual grounding on a city-scale 3D point cloud.

3174. Not All Language Model Features Are One-Dimensionally Linear

链接：https://iclr.cc/virtual/2025/poster/29008 abstract： Recent work has proposed that language models perform computation by manipulating one-dimensional representations of concepts ("features") in activation space. In contrast, we explore whether some language model representations may be inherently multi-dimensional. We begin by developing a rigorous definition of irreducible multi-dimensional features based on whether they can be decomposed into either independent or non-co-occurring lower-dimensional features. Motivated by these definitions, we design a scalable method that uses sparse autoencoders to automatically find multi-dimensional features in GPT-2 and Mistral 7B. These auto-discovered features include strikingly interpretable examples, e.g. $\textit{circular}$ features representing days of the week and months of the year. We identify tasks where these exact circles are used to solve computational problems involving modular arithmetic in days of the week and months of the year. Next, we provide evidence that these circular features are indeed the fundamental unit of computation in these tasks with intervention experiments on Mistral 7B and Llama 3 8B, and we examine the continuity of the days of the week feature in Mistral 7B. Overall, our work argues that understanding multi-dimensional features is necessary to mechanistically decompose some model behaviors.

3175. MVTokenFlow: High-quality 4D Content Generation using Multiview Token Flow

链接：https://iclr.cc/virtual/2025/poster/27643 abstract： In this paper, we present MVTokenFlow for high-quality 4D content creation from monocular videos. Recent advancements in generative models such as video diffusion models and multiview diffusion models enable us to create videos or 3D models. However, extending these generative models for dynamic 4D content creation is still a challenging task that requires the generated content to be consistent spatially and temporally. To address this challenge, MVTokenFlow utilizes the multiview diffusion model to generate multiview images on different timesteps, which attains spatial consistency across different viewpoints and allows us to reconstruct a reasonable coarse 4D field. Then, MVTokenFlow further regenerates all the multiview images using the rendered 2D flows as guidance. The 2D flows effectively associate pixels from different timesteps and improve the temporal consistency by reusing tokens in the regeneration process. Finally, the regenerated images are spatiotemporally consistent and utilized to refine the coarse 4D field to get a high-quality 4D field. Experiments demonstrate the effectiveness of our design and show significantly improved quality than baseline methods. Project page: https://soolab.github.io/MVTokenFlow.

3176. Rapidly Adapting Policies to the Real-World via Simulation-Guided Fine-Tuning

链接：https://iclr.cc/virtual/2025/poster/29282 abstract： Robot learning requires a considerable amount of high-quality data to realize the promise of generalization. However, large data sets are costly to collect in the real world. Physics simulators can cheaply generate vast data sets with broad coverage over states, actions, and environments. However, physics engines are

fundamentally misspecified approximations to reality. This makes direct zero-shot transfer from simulation to reality challenging, especially in tasks where precise and force-sensitive manipulation is necessary. Thus, fine-tuning these policies with small real-world data sets is an appealing pathway for scaling robot learning. However, current reinforcement learning fine-tuning frameworks leverage general, unstructured exploration strategies which are too inefficient to make real-world adaptation practical. This paper introduces the \emph{Simulation-Guided Fine-tuning} (SGFT) framework, which demonstrates how to extract structural priors from physics simulators to substantially accelerate real-world adaptation. Specifically, our approach uses a value function learned in simulation to guide real-world exploration. We demonstrate this approach across five real-world dexterous manipulation tasks where zero-shot sim-to-real transfer fails. We further demonstrate our framework substantially outperforms baseline fine-tuning methods, requiring up to an order of magnitude fewer real-world samples and succeeding at difficult tasks where prior approaches fail entirely. Last but not least, we provide theoretical justification for this new paradigm which underpins how SGFT can rapidly learn high-performance policies in the face of large sim-to-real dynamics gaps.

3177. Modality-Specialized Synergizers for Interleaved Vision-Language Generalists

链接：https://iclr.cc/virtual/2025/poster/30822 abstract： Recent advancements in Vision-Language Models (VLMs) have led to the emergence of Vision-Language Generalists (VLGs) capable of understanding and generating both text and images. However, seamlessly generating an arbitrary sequence of text and images remains a challenging task for the current VLGs. One primary limitation lies in applying a unified architecture and the same set of parameters to simultaneously model discrete text tokens and continuous image features. Recent works attempt to tackle this fundamental problem by introducing modality-aware expert models. However, they employ identical architectures to process both text and images, disregarding the intrinsic inductive biases in these two modalities. In this work, we introduce Modality-Specialized Synergizers (MoSS), a novel design that efficiently optimizes existing unified architectures of VLGs with modality-specialized adaptation layers, i.e., a Convolutional LoRA for modeling the local priors of image patches and a Linear LoRA for processing sequential text. This design enables more effective modeling of modality-specific features while maintaining the strong cross-modal integration gained from pretraining. In addition, to improve the instruction-following capability on interleaved text-and-image generation, we introduce LeafInstruct, the first open-sourced interleaved instruction tuning dataset comprising 184,982 high-quality instances on more than 10 diverse domains. Extensive experiments show that VLGs integrated with MoSS achieve state-of-the-art performance, significantly surpassing baseline VLGs in complex interleaved generation tasks. Furthermore, our method exhibits strong generalizability on different VLGs.

3178. SPARTUN3D: Situated Spatial Understanding of 3D World in Large Language Model

链接：https://iclr.cc/virtual/2025/poster/30351 abstract： Integrating the 3D world into large language models (3D-based LLMs) has been a promising research direction for 3D scene understanding. However, current 3D-based LLMs fall short in situated understanding due to two key limitations: 1) existing 3D datasets are constructed from a global perspective of the 3D scenes and lack situated context.2) the architectures of the current 3D-based LLMs lack an explicit mechanism for aligning situated spatial information between 3D representations and natural language, limiting their performance in tasks requiring precise spatial reasoning. In this work, we address these issues by introducing a scalable situated 3D dataset, named Spartun3D, that incorporates various situated spatial information.In addition, we propose a situated spatial alignment module to enhance the learning between 3D visual representations and their corresponding textual descriptions. Our experimental results demonstrate that both our dataset and alignment module enhance situated spatial understanding ability.

3179. State Space Models are Provably Comparable to Transformers in Dynamic Token Selection

链接：https://iclr.cc/virtual/2025/poster/29700 abstract： Deep neural networks based on state space models (SSMs) are attracting significant attention in sequence modeling since their computational cost is much smaller than that of Transformers. While the capabilities of SSMs have been demonstrated through experiments in various tasks, theoretical understanding of SSMs is still limited. In particular, most theoretical studies discuss the capabilities of SSM layers without nonlinear layers, and there is a lack of discussion on their combination with nonlinear layers. In this paper, we explore the capabilities of SSMs combined with fully connected neural networks, and show that they are comparable to Transformers in extracting the essential tokens depending on the input. As concrete examples, we consider two synthetic tasks, which are challenging for a single SSM layer, and demonstrate that SSMs combined with nonlinear layers can efficiently solve these tasks. Furthermore, we study the nonparametric regression task, and prove that the ability of SSMs is equivalent to that of Transformers in estimating functions belonging to a certain class.

3180. AgentStudio: A Toolkit for Building General Virtual Agents

链接：https://iclr.cc/virtual/2025/poster/29133 abstract： General virtual agents need to handle multimodal observations, master complex action spaces, and self-improve in dynamic, open-domain environments. However, existing environments are often domain-specific and require complex setups, which limits agent development and evaluation in real-world settings. As a result, current evaluations lack in-depth analyses that decompose fundamental agent capabilities. We introduce AgentStudio, a trinity

of environments, tools, and benchmarks to address these issues. AgentStudio provides a lightweight, interactive environment with highly generic observation and action spaces, e.g., video observations and GUI/API actions. It integrates tools for creating online benchmark tasks, annotating GUI elements, and labeling actions in videos. Based on our environment and tools, we curate an online task suite that benchmarks both GUI interactions and function calling with efficient auto-evaluation. We also reorganize existing datasets and collect new ones using our tools to establish three datasets: GroundUI, IDMBench, and CriticBench. These datasets evaluate fundamental agent abilities, including GUI grounding, learning from videos, and success detection, pointing to the desiderata for robust, general, and open-ended virtual agents.

3181. Unleashing the Power of Task-Specific Directions in Parameter Efficient Fine-tuning

链接：https://iclr.cc/virtual/2025/poster/32083 abstract： Large language models demonstrate impressive performance on downstream tasks, yet requiring extensive resource consumption when fully fine-tuning all parameters. To mitigate this, Parameter Efficient Fine-Tuning (PEFT) strategies, such as LoRA, have been developed. In this paper, we delve into the concept of task-specific directions (TSDs)—critical for transitioning large models from pretrained states to task-specific enhancements in PEFT. We propose a framework to clearly define these directions and explore their properties, and practical utilization challenges. We then introduce a novel approach, LoRA-Dash, which aims to maximize the impact of TSDs during the fine-tuning process, thereby enhancing model performance on targeted tasks. Extensive experiments have conclusively demonstrated the effectiveness of LoRA-Dash, and in-depth analyses further reveal the underlying mechanisms of LoRA-Dash.

3182. Diffusing to the Top: Boost Graph Neural Networks with Minimal Hyperparameter Tuning

链接：https://iclr.cc/virtual/2025/poster/30476 abstract： Graph Neural Networks (GNNs) are proficient in graph representation learning and achieve promising performance on versatile tasks such as node classification and link prediction.Usually, a comprehensive hyperparameter tuning is essential for fully unlocking GNN's top performance, especially for complicated tasks such as node classification on large graphs and long-range graphs. This is usually associated with high computational and time costs and careful design of appropriate search spaces. This work introduces a graph-conditioned latent diffusion framework (GNN-Diff) to generate high-performing GNNs based on the model checkpoints of sub-optimal hyperparameters selected by a light-tuning coarse search. We validate our method through 166 experiments across four graph tasks: node classification on small, large, and long-range graphs, as well as link prediction. Our experiments involve 10 classic and state-of-the-art target models and 20 publicly available datasets. The results consistently demonstrate that GNN-Diff: (1) boosts the performance of GNNs with efficient hyperparameter tuning; and (2) presents high stability and generalizability on unseen data across multiple generation runs. The code is available at https://github.com/lequanlin/GNN-Diff.

3183. When Graph Neural Networks Meet Dynamic Mode Decomposition

链接：https://iclr.cc/virtual/2025/poster/28958 abstract： Graph Neural Networks (GNNs) have emerged as fundamental tools for a wide range of prediction tasks on graph-structured data. Recent studies have drawn analogies between GNN feature propagation and diffusion processes, which can be interpreted as dynamical systems. In this paper, we delve deeper into this perspective by connecting the dynamics in GNNs to modern Koopman theory and its numerical method, Dynamic Mode Decomposition (DMD). We illustrate how DMD can estimate a low-rank, finite-dimensional linear operator based on multiple states of the system, effectively approximating potential nonlinear interactions between nodes in the graph. This approach allows us to capture complex dynamics within the graph accurately and efficiently. We theoretically establish a connection between the DMD-estimated operator and the original dynamic operator between system states. Building upon this foundation, we introduce a family of DMD-GNN models that effectively leverage the low-rank eigenfunctions provided by the DMD algorithm. We further discuss the potential of enhancing our approach by incorporating domain-specific constraints such as symmetry into the DMD computation, allowing the corresponding GNN models to respect known physical properties of the underlying system. Our work paves the path for applying advanced dynamical system analysis tools via GNNs. We validate our approach through extensive experiments on various learning tasks, including directed graphs, large-scale graphs, long-range interactions, and spatial-temporal graphs. We also empirically verify that our proposed models can serve as powerful encoders for link prediction tasks. The results demonstrate that our DMD-enhanced GNNs achieve state-of-the-art performance, highlighting the effectiveness of integrating DMD into GNN frameworks.

3184. Draw-and-Understand: Leveraging Visual Prompts to Enable MLLMs to Comprehend What You Want

链接：https://iclr.cc/virtual/2025/poster/29098 abstract： In this paper, we present the Draw-and-Understand framework, exploring how to integrate visual prompting understanding capabilities into Multimodal Large Language Models (MLLMs). Visual prompts allow users to interact through multi-modal instructions, enhancing the models' interactivity and fine-grained image comprehension. In this framework, we propose a general architecture adaptable to different pre-trained MLLMs, enabling it to recognize various types of visual prompts (such as points, bounding boxes, and free-form shapes) alongside language understanding. Additionally, we introduce MDVP-Instruct-Data, a multi-domain dataset featuring 1.2 million image-visual prompt-text triplets, including natural images, document images, scene text images, mobile/web screenshots, and remote sensing

images. Building on this dataset, we introduce MDVP-Bench, a challenging benchmark designed to evaluate a model's ability to understand visual prompting instructions. The experimental results demonstrate that our framework can be easily and effectively applied to various MLLMs, such as SPHINX-X and LLaVA. After training with MDVP-Instruct-Data and image-level instruction datasets, our models exhibit impressive multimodal interaction capabilities and pixel-level understanding, while maintaining their image-level visual perception performance.

3185. Linear Transformer Topological Masking with Graph Random Features

链接：https://iclr.cc/virtual/2025/poster/30895 abstract： When training transformers on graph-structured data, incorporating information about the underlying topology is crucial for good performance. Topological masking, a type of relative position encoding, achieves this by upweighting or downweighting attention depending on the relationship between the query and keys in the graph. In this paper, we propose to parameterise topological masks as a learnable function of a weighted adjacency matrix -- a novel, flexible approach which incorporates a strong structural inductive bias. By approximating this mask with graph random features (for which we prove the first known concentration bounds), we show how this can be made fully compatible with linear attention, preserving $\mathcal{O}(N)$ time and space complexity with respect to the number of input tokens. The fastest previous alternative was $\mathcal{O}(N \log N)$ and only suitable for specific graphs. Our efficient masking algorithms provide strong performance gains for image and point cloud data, including with >30k nodes.

3186. Model-based RL as a Minimalist Approach to Horizon-Free and Second-Order Bounds

链接：https://iclr.cc/virtual/2025/poster/28008 abstract： Learning a transition model via Maximum Likelihood Estimation (MLE) followed by planning inside the learned model is perhaps the most standard and simplest Model-based Reinforcement Learning (RL) framework. In this work, we show that such a simple Model-based RL scheme, when equipped with optimistic and pessimistic planning procedures, achieves strong regret and sample complexity bounds in online and offline RL settings. Particularly, we demonstrate that under the conditions where the trajectory-wise reward is normalized between zero and one and the transition is time-homogenous, it achieves nearly horizon-free and second-order bounds.

3187. DINOv2: Learning Robust Visual Features without Supervision

链接：https://iclr.cc/virtual/2025/poster/31510 abstract： The recent breakthroughs in natural language processing for model pretraining on large quantities of data have opened the way for similar foundation models in computer vision. These models could greatly simplify the use of images in any system by producing all-purpose visual features, i.e., features that work across image distributions and tasks without finetuning. This work shows that existing pretraining methods, especially self-supervised methods, can produce such features if trained on enough curated data from diverse sources. We revisit existing approaches and combine different techniques to scale our pretraining in terms of data and model size. Most of the technical contributions aim at accelerating and stabilizing the training at scale. In terms of data, we propose an automatic pipeline to build a dedicated, diverse, and curated image dataset instead of uncurated data, as typically done in the self-supervised literature. In terms of models, we train a ViT model with 1B parameters and distill it into a series of smaller models that surpass the best available all-purpose features, OpenCLIP on most of the benchmarks at image and pixel levels.

3188. Learning under Temporal Label Noise

链接：https://iclr.cc/virtual/2025/poster/30932 abstract： Many time series classification tasks, where labels vary over time, are affected by label noise that also varies over time. Such noise can cause label quality to improve, worsen, or periodically change over time. We first propose and formalize temporal label noise, an unstudied problem for sequential classification of time series. In this setting, multiple labels are recorded over time while being corrupted by a time-dependent noise function. We first demonstrate the importance of modeling the temporal nature of the label noise function and how existing methods will consistently underperform. We then propose methods to train noise-tolerant classifiers by estimating the temporal label noise function directly from data. We show that our methods lead to state-of-the-art performance under diverse types of temporal label noise on real-world datasets.

3189. Robustness of Quantum Algorithms for Nonconvex Optimization

链接：https://iclr.cc/virtual/2025/poster/30083 abstract： In this paper, we systematically study quantum algorithms for finding an ϵ-approximate second-order stationary point (ϵ-SOSP) of a d-dimensional nonconvex function, a fundamental problem in nonconvex optimization, with noisy zeroth- or first-order oracles as inputs. We first prove that, up to noise of $O(\epsilon^{10}/d^5)$, perturbed accelerated gradient descent equipped with quantum gradient estimation takes $O(\log d/\epsilon^{1.75})$ quantum queries to find an ϵ-SOSP. We then prove that standard perturbed gradient descent is robust to the noise of $O(\epsilon^6/d^4)$ and $O(\epsilon/d^{0.5+\zeta})$ for any $\zeta>0$ on the zeroth- and first-order oracles, respectively, which provides a quantum algorithm with poly-logarithmic query complexity. Furthermore, we propose a stochastic gradient descent algorithm using quantum mean estimation on the Gaussian smoothing of noisy oracles, which is robust to $O(\epsilon^{1.5}/d)$ and $O(\epsilon/\sqrt{d})$ noise on the zeroth- and first-order oracles, respectively. The quantum algorithm takes $O(d^{2.5}/\epsilon^{3.5})$ and $O(d^2/\epsilon^3)$ queries to the two oracles, giving a polynomial speedup over the classical counterparts. As a complement, we characterize the domains where quantum algorithms can find an

ϵ-SOSP with poly-logarithmic, polynomial, or exponential number of queries in d, or the problem is information-theoretically unsolvable even with an infinite number of queries. In addition, we prove an $\Omega(\epsilon^{-12/7})$ lower bound on ϵ for any randomized classical and quantum algorithm to find an ϵ-SOSP using either noisy zeroth- or first-order oracles.

3190. Uncertainty Herding: One Active Learning Method for All Label Budgets

链接：https://iclr.cc/virtual/2025/poster/29463 abstract： Most active learning research has focused on methods which perform well when many labels are available, but can be dramatically worse than random selection when label budgets are small.Other methods have focused on the low-budget regime, but do poorly as label budgets increase.As the line between "low" and "high" budgets varies by problem,this is a serious issue in practice.We propose uncertainty coverage,an objective which generalizes a variety of low- and high-budget objectives,as well as natural, hyperparameter-light methods to smoothly interpolate between low- and high-budget regimes.We call greedy optimization of the estimate Uncertainty Herding;this simple method is computationally fast,and we prove that it nearly optimizes the distribution-level coverage.In experimental validation across a variety of active learning tasks,our proposal matches or beats state-of-the-art performance in essentially all cases;it is the only method of which we are aware that reliably works well in both low- and high-budget settings.

3191. What Has Been Overlooked in Contrastive Source-Free Domain Adaptation: Leveraging Source-Informed Latent Augmentation within Neighborhood Context

链接：https://iclr.cc/virtual/2025/poster/31489 abstract： Source-free domain adaptation (SFDA) involves adapting a model originally trained using a labeled dataset (source domain) to perform effectively on an unlabeled dataset (target domain) without relying on any source data during adaptation. This adaptation is especially crucial when significant disparities in data distributions exist between the two domains and when there are privacy concerns regarding the source model's training data. The absence of access to source data during adaptation makes it challenging to analytically estimate the domain gap. To tackle this issue, various techniques have been proposed, such as unsupervised clustering, contrastive learning, and continual learning. In this paper, we first conduct an extensive theoretical analysis of SFDA based on contrastive learning, primarily because it has demonstrated superior performance compared to other techniques. Motivated by the obtained insights, we then introduce a straightforward yet highly effective latent augmentation method tailored for contrastive SFDA. This augmentation method leverages the dispersion of latent features within the neighborhood of the query sample, guided by the source pre-trained model, to enhance the informativeness of positive keys. Our approach, based on a single InfoNCE-based contrastive loss, outperforms state-of-the-art SFDA methods on widely recognized benchmark datasets.

3192. Reducing Hallucinations in Large Vision-Language Models via Latent Space Steering

链接：https://iclr.cc/virtual/2025/poster/30013 abstract： Hallucination poses a challenge to the deployment of large vision-language models (LVLMs) in applications. Unlike in large language models (LLMs), hallucination in LVLMs often arises from misalignments between visual inputs and textual outputs. This paper investigates the underlying mechanisms of hallucination, focusing on the unique structure of LVLMs that distinguishes them from LLMs. We identify that hallucinations often arise from the sensitivity of text decoders to vision inputs, a natural phenomenon when image encoders and text decoders are pre-trained separately. Inspired by this, we introduce Visual and Textual Intervention (VTI), a novel technique designed to reduce hallucinations by steering latent space representations during inference to enhance the stability of vision features. As a task-agnostic test-time intervention, VTI can be easily applied to any problem without additional training costs. Extensive experiments demonstrate that it can effectively reduce hallucinations and outperform baseline methods across multiple metrics, highlighting the critical role of vision feature stability in LVLMs.

3193. TorchTitan: One-stop PyTorch native solution for production ready LLM pretraining

链接：https://iclr.cc/virtual/2025/poster/29620 abstract： The development of large language models (LLMs) has been instrumental in advancing state-of-the-art natural language processing applications. Training LLMs with billions of parameters and trillions of tokens requires sophisticated distributed systems that enable composing and comparing several state-of-the-art techniques in order to efficiently scale across thousands of accelerators. However, existing solutions are complex, scattered across multiple libraries/repositories, lack interoperability, and are cumbersome to maintain. Thus, curating and empirically comparing training recipes requires non-trivial engineering effort.This paper introduces **TORCHTITAN**^{1}, a PyTorch-native distributed training system that unifies and advances state-of-the-art techniques, streamlining integration and reducing engineering overhead. TORCHTITAN enables seamless application of 4D parallelism in a modular and composable manner, while featuring elastic scaling to adapt to changing computational requirements. The system provides comprehensive logging, efficient checkpointing, and debugging tools, ensuring production-ready training. Moreover, TORCHTITAN incorporates innovative hardware-software co-designed solutions, leveraging cutting-edge features like Float8 training and

SymmetricMemory to maximize hardware utilization.As a flexible experimental test bed, TORCHTITAN facilitates the curation and comparison of custom recipes for diverse training contexts. By leveraging TORCHTITAN, we developed optimized training recipes for the Llama 3.1 family and provide actionable guidance on selecting and combining distributed training techniques to maximize training efficiency, based on our hands-on experiences.We thoroughly assess TORCHTITAN on the Llama 3.1 family of LLMs, spanning 8 billion to 405 billion parameters, and showcase its exceptional performance, modular composability, and elastic scalability. By stacking training optimizations, we demonstrate accelerations ranging from 65.08% on Llama 3.1 8B at 128 GPU scale (1D), 12.59% on Llama 3.1 70B at 256 GPU scale (2D), to 30% on Llama 3.1 405B at 512 GPU scale (3D) on NVIDIA H100 GPUs over optimized baselines. We also demonstrate the effectiveness of 4D parallelism in enabling long context training.^{1} GitHub: https://github.com/pytorch/torchtitan

3194. HELM: Hierarchical Encoding for mRNA Language Modeling

链接：https://iclr.cc/virtual/2025/poster/29936 abstract： Messenger RNA (mRNA) plays a crucial role in protein synthesis, with its codon structure directly impacting biological properties. While Language Models (LMs) have shown promise in analyzing biological sequences, existing approaches fail to account for the hierarchical nature of mRNA's codon structure. We introduce Hierarchical Encoding for mRNA Language Modeling (HELM), a novel pre-training strategy that incorporates codon-level hierarchical structure into language model training. HELM modulates the loss function based on codon synonymity, aligning the model's learning process with the biological reality of mRNA sequences. We evaluate HELM on diverse mRNA datasets and tasks, demonstrating that HELM outperforms standard language model pre-training as well as existing foundation model baselines on six diverse downstream property prediction tasks and an antibody region annotation tasks on average by around 8%. Additionally, HELM enhances the generative capabilities of language model, producing diverse mRNA sequences that better align with the underlying true data distribution compared to non-hierarchical baselines.

3195. ShortcutsBench: A Large-Scale Real-world Benchmark for API-based Agents

链接：https://iclr.cc/virtual/2025/poster/28594 abstract： Recent advancements in integrating large language models (LLMs) with application programming interfaces (APIs) have gained significant interest in both academia and industry. Recent work demonstrates that these API-based agents exhibit relatively strong autonomy and planning capabilities. However, their ability to handle multi-dimensional difficulty levels, diverse task types, and real-world demands remains unknown. In this paper, we introduce ShortcutsBench, a large-scale benchmark for the comprehensive evaluation of API-based agents in solving real-world complex tasks. ShortcutsBench includes a wealth of real APIs from Apple Inc., refined user queries, human-annotated high-quality action sequences, detailed parameter filling values, and parameters requesting necessary input from the system or user. We put in significant effort in collecting and processing the data. We revealed how existing benchmarks / datasets struggle to accommodate the advanced reasoning capabilities of existing more intelligent LLMs. Moreover, our extensive evaluation of agents built with 5 leading open-source (size >= 57B) and 5 closed-source LLMs (e.g. Gemini-1.5-Pro and GPT-4o-mini) reveals significant limitations of existing API-based agents in the whole process of handling complex queries related to API selection, parameter filling, and requesting necessary input from the system and the user. These findings highlight the great challenges that API-based agents face in effectively fulfilling real and complex user queries. All datasets, code, experimental logs, and results are available at \url{https://anonymous.4open.science/r/ShortcutsBench}.

3196. q-exponential family for policy optimization

链接：https://iclr.cc/virtual/2025/poster/29786 abstract： Policy optimization methods benefit from a simple and tractable policy parametrization, usually the Gaussian for continuous action spaces. In this paper, we consider a broader policy family that remains tractable: the q-exponential family. This family of policies is flexible, allowing the specification of both heavy-tailed policies ($q>1$) and light-tailed policies ($q<1$). This paper examines the interplay between q-exponential policies for several actor-critic algorithms conducted on both online and offline problems. We find that heavy-tailed policies are more effective in general and can consistently improve on Gaussian. In particular, we find the Student's t-distribution to be more stable than the Gaussian across settings and that a heavy-tailed q-Gaussian for Tsallis Advantage Weighted Actor-Critic consistently performs well in offline benchmark problems.In summary, we find that the Student's t policy a strong candidate for drop-in replacement to the Gaussian.Our code is available at \url{https://github.com/lingweizhu/qexp}.

3197. SEAL: Safety-enhanced Aligned LLM Fine-tuning via Bilevel Data Selection

链接：https://iclr.cc/virtual/2025/poster/29422 abstract： Fine-tuning on task-specific data to boost downstream performance is a crucial step for leveraging Large Language Models (LLMs). However, though fine-tuning enhances the model performance for specialized applications, previous studies have demonstrated that fine-tuning the models on several adversarial samples or even benign data can greatly comprise the model's pre-equipped alignment and safety capabilities. In this work, we propose SEAL, a novel framework to enhance safety in LLM fine-tuning. SEAL learns a data ranker based on the bilevel optimization to up rank the safe and high-quality fine-tuning data and down rank the unsafe or low-quality ones. Models trained with SEAL demonstrate superior quality over multiple baselines, with 8.5\% and 9.7\% win rate increase compared to random selection respectively on Llama-3-8b-Instruct and Merlinite-7b models. Our code is available on github https://github.com/hanshen95/SEAL.

3198. Do as We Do, Not as You Think: the Conformity of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28094 abstract： Recent advancements in large language models (LLMs) revolutionize the field of intelligent agents, enabling collaborative multi-agent systems capable of tackling complex problems across various domains. However, the potential of conformity within these systems, analogous to phenomena like conformity bias and group-think in human group dynamics, remains largely unexplored, raising concerns about their collective problem-solving capabilities and possible ethical implications. This paper presents a comprehensive study on conformity in LLM-driven multi-agent systems, focusing on three aspects: the existence of conformity, the factors influencing conformity, and potential mitigation strategies. In particular, we introduce BenchForm, a new conformity-oriented benchmark, featuring reasoning-intensive tasks and five distinct interaction protocols designed to probe LLMs' behavior in collaborative scenarios. Several representative LLMs are evaluated on BenchForm, using metrics such as conformity rate and independence rate to quantify conformity's impact. Our analysis delves into factors influencing conformity, including interaction time and majority size, and examines how the subject agent rationalize its conforming behavior. Furthermore, we explore two strategies to mitigate conformity effects, i.e., developing enhanced persona and implementing a reflection mechanism. Several interesting findings regarding LLMs' conformity are derived from empirical results and case studies. We hope that these insights can pave the way for more robust and ethically-aligned collaborative AI systems. Our benchmark and code are available at BenchForm.

3199. Addressing Label Shift in Distributed Learning via Entropy Regularization

链接：https://iclr.cc/virtual/2025/poster/28558 abstract： We address the challenge of minimizing "true risk" in multi-node distributed learning.\footnote{We use the term node to refer to a client, FPGA, APU, CPU, GPU, or worker.} These systems are frequently exposed to both inter-node and intra-node "label shifts", which present a critical obstacle to effectively optimizing model performance while ensuring that data remains confined to each node.To tackle this, we propose the Versatile Robust Label Shift (VRLS) method, which enhances the maximum likelihood estimation of the test-to-train label importance ratio. VRLS incorporates Shannon entropy-based regularization and adjusts the importance ratio during training to better handle label shifts at the test time.In multi-node learning environments, VRLS further extends its capabilities by learning and adapting importance ratios across nodes, effectively mitigating label shifts and improving overall model performance. Experiments conducted on MNIST, Fashion MNIST, and CIFAR-10 demonstrate the effectiveness of VRLS, outperforming baselines by up to 20\% in imbalanced settings. These results highlight the significant improvements VRLS offers in addressing label shifts. Our theoretical analysis further supports this by establishing high-probability bounds on estimation errors.

3200. LIFe-GoM: Generalizable Human Rendering with Learned Iterative Feedback Over Multi-Resolution Gaussians-on-Mesh

链接：https://iclr.cc/virtual/2025/poster/28812 abstract： Generalizable rendering of an animatable human avatar from sparse inputs relies on data priors and inductive biases extracted from training on large data to avoid scene-specific optimization and to enable fast reconstruction. This raises two main challenges: First, unlike iterative gradient-based adjustment in scene-specific optimization, generalizable methods must reconstruct the human shape representation in a single pass at inference time.Second, rendering is preferably computationally efficient yet of high resolution.To address both challenges we augment the recently proposed dual shape representation, which combines the benefits of a mesh and Gaussian points, in two ways. To improve reconstruction, we propose an iterative feedback update framework, which successively improves the canonical human shape representation during reconstruction.To achieve computationally efficient yet high-resolution rendering, we study a coupled-multi-resolution Gaussians-on-Mesh representation.We evaluate the proposed approach on the challenging THuman2.0, XHuman and AIST++ data. Our approach reconstructs an animatable representation from sparse inputs in less than 1s, renders views with 95.1FPS at 1024×1024, and achieves PSNR/LPIPS*/FID of 24.65/110.82/51.27 on THuman2.0, outperforming the state-of-the-art in rendering quality.