# 2401. Efficient Low-Bit Quantization with Adaptive Scales for Multi-Task Co-Training

链接：https://iclr.cc/virtual/2025/poster/27868 abstract： Co-training can achieve parameter-efficient multi-task models but remains unexplored for quantization-aware training. Our investigation shows that directly introducing co-training into existing quantization-aware training (QAT) methods results in significant performance degradation. Our experimental study identifies that the primary issue with existing QAT methods stems from the inadequate activation quantization scales for the co-training framework. To address this issue, we propose Task-Specific Scales Quantization for Multi-Task Co-Training (TSQ-MTC) to tackle mismatched quantization scales. Specifically, a task-specific learnable multi-scale activation quantizer (TLMAQ) is incorporated to enrich the representational ability of shared features for different tasks. Additionally, we find that in the deeper layers of the Transformer model, the quantized network suffers from information distortion within the attention quantizer. A structure-based layer-by-layer distillation (SLLD) is then introduced to ensure that the quantized features effectively preserve the information from their full-precision counterparts. Our extensive experiments in two co-training scenarios demonstrate the effectiveness and versatility of TSQ-MTC. In particular, we successfully achieve a 4-bit quantized low-level visual foundation model based on IPT, which attains a PSNR comparable to the full-precision model while offering a $7.99\times$ compression ratio in the $\times4$ super-resolution task on the Set5 benchmark.

# 2402. SelKD: Selective Knowledge Distillation via Optimal Transport Perspective

链接：https://iclr.cc/virtual/2025/poster/30244 abstract： Knowledge Distillation (KD) has been a popular paradigm for training a (smaller) student model from its teacher model. However, little research has been done on the practical scenario where only a subset of the teacher's knowledge needs to be distilled, which we term selective KD (SelKD). This demand is especially pronounced in the era of foundation models, where the teacher model can be significantly larger than the student model. To address this issue, we propose to rethink the knowledge distillation problem from the perspective of Inverse Optimal Transport (IOT). Previous Bayesian frameworks mapped each sample to the probabilities of corresponding labels in an end-to-end manner, which fixed the number of classification categories and hindered effective partial knowledge transfer. In contrast, IOT calculates from the standpoint of transportation or matching, allowing for the flexible selection of samples and their quantities for matching. Traditional logit-based KD can be viewed as a special case within the IOT framework. Building on this IOT foundation, we formalize this setting in the context of classification, where only selected categories from the teacher's category space are required to be recognized by the student in the context of closed-set recognition, which we call closed-set SelKD, enhancing the student's performance on specific subtasks. Furthermore, we extend the closed-set SelKD, introducing an open-set version of SelKD, where the student model is required to provide a "not selected" response for categories outside its assigned task. Experimental results on standard benchmarks demonstrate the superiority of our approach. The source code is available at: \href{https://github.com/machoshi/SelKD}

# 2403. Regularizing Energy among Training Samples for Out-of-Distribution Generalization

链接：https://iclr.cc/virtual/2025/poster/29985 abstract：

# 2404. Rethinking and Improving Autoformalization: Towards a Faithful Metric and a Dependency Retrieval-based Approach

链接：https://iclr.cc/virtual/2025/poster/28759 abstract： As a central component in formal verification, statement autoformalization has been widely studied including the recent efforts from machine learning community, but still remains a widely-recognized difficult and open problem. In this paper, we delve into two critical yet under-explored gaps: 1) absence of faithful and universal automated evaluation for autoformalization results; 2) agnosia of contextual information, inducing severe hallucination of formal definitions and theorems. To address the first issue, we propose **BEq** (***B**idirectional **E**xtended Definitional E**q**uivalence*), an automated neuro-symbolic method to determine the equivalence between two formal statements, which is formal-grounded and well-aligned with human intuition. For the second, we propose **RAutoformalizer** (***R**etrieval-augmented **Autoformalizer***), augmenting statement autoformalization by *Dependency Retrieval*, retrieving potentially dependent objects from formal libraries. We parse the dependencies of libraries and propose to *structurally informalise* formal objects by the topological order of dependencies. To evaluate OOD generalization and research-level capabilities, we build a novel benchmark, *Con-NF*, consisting of 961 informal-formal statement pairs from frontier mathematical researches. Experiments validate the effectiveness of our approaches: BEq is evaluated on 200 diverse formal statement pairs with expert-annotated equivalence label, exhibiting significantly improved accuracy ($82.50\% \mapsto 90.50\%$) and precision ($70.59\% \mapsto 100.0\%$). For dependency retrieval, a strong baseline is devised. Our RAutoformalizer substantially outperforms SOTA baselines in both in-distribution ProofNet benchmark ($12.83\% \mapsto 18.18\%$, BEq@8) and OOD Con-NF scenario ($4.58\%\mapsto 16.86\%$, BEq@8).

# 2405. Learning Structured Universe Graph with Outlier OOD Detection for

# Partial Matching

链接：https://iclr.cc/virtual/2025/poster/32072 abstract： Partial matching is a kind of graph matching where only part of two graphs can be aligned. This problem is particularly important in computer vision applications, where challenges like point occlusion or annotation errors often occur when labeling key points. Previous work has often conflated point occlusion and annotation errors, despite their distinct underlying causes. We propose two components to address these challenges: (1) a structured universe graph is learned to connect two input graphs $X_{ij} = X_{iu} X_{ju}^\top$, effectively resolving the issue of point occlusion; (2) an energy-based out-of-distribution detection is designed to remove annotation errors from the input graphs before matching. We evaluated our method on the Pascal VOC and Willow Object datasets, focusing on scenarios involving point occlusion and random outliers. The experimental results demonstrate that our approach consistently outperforms state-of-the-art methods across all tested scenarios, highlighting the accuracy and robustness of our method.

# 2406. What Secrets Do Your Manifolds Hold? Understanding the Local Geometry of Generative Models

链接：https://iclr.cc/virtual/2025/poster/28903 abstract： Deep Generative Models are frequently used to learn continuous representations of complex data distributions by training on a finite number of samples. For any generative model, including pre-trained foundation models with Diffusion or Transformer architectures, generation performance can significantly vary across the learned data manifold. In this paper, we study the local geometry of the learned manifold and its relationship to generation outcomes for a wide range of generative models, including DDPM, Diffusion Transformer (DiT), and Stable Diffusion 1.4. Building on the theory of continuous piecewise-linear (CPWL) generators, we characterize the local geometry in terms of three geometric descriptors - scaling ($\psi$), rank ($\nu$), and complexity/un-smoothness ($\delta$). We provide quantitative and qualitative evidence showing that for a given latent vector, the local descriptors are indicative of post-generation aesthetics, generation diversity, and memorization by the generative model. Finally, we demonstrate that by training a reward model on the 'local scaling' for Stable Diffusion, we can self-improve both generation aesthetics and diversity using geometry sensitive guidance during denoising. Website: https://imtiazhumayun.github.io/generative_geometry.

# 2407. To Clip or not to Clip: the Dynamics of SGD with Gradient Clipping in High-Dimensions

链接：https://iclr.cc/virtual/2025/poster/28617 abstract： The success of modern machine learning is due in part to the adaptive optimization methods that have been developed to deal with the difficulties of training large models over complex datasets. One such method is gradient clipping: a practical procedure with limited theoretical underpinnings. In this work, we study clipping in a least squares problem under streaming SGD. We develop a theoretical analysis of the learning dynamics in the limit of large intrinsic dimension—a model and dataset dependent notion of dimensionality. In this limit we find a deterministic equation that describes the evolution of the loss and demonstrate that this equation predicts the path of clipped SGD on synthetic, CIFAR10, and Wikitext2 data. We show that with Gaussian noise clipping cannot improve SGD performance. Yet, in other noisy settings, clipping can provide benefits with tuning of the clipping threshold. We propose a simple heuristic for near optimal scheduling of the clipping threshold which requires the tuning of only one hyperparameter. We conclude with a discussion about the links between high-dimensional clipping and neural network training.

# 2408. Locality Alignment Improves Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/28233 abstract： Vision language models (VLMs) have seen growing adoption in recent years, but many still struggle with basic spatial reasoning errors. We hypothesize that this is due to VLMs adopting pre-trained vision backbones, specifically vision transformers (ViTs) trained with image-level supervision and minimal inductive biases. Such models may fail to encode the class contents at each position in the image, and our goal is to resolve this with a vision backbone that effectively captures both local and global image semantics. Our main insight is that we do not require new supervision to learn this capability – pre-trained models contain significant knowledge of local semantics that we can extract and use for scalable self-supervision. We propose a new efficient post-training stage for ViTs called locality alignment and a novel fine-tuning procedure called MaskEmbed that uses a masked reconstruction loss to learn semantic contributions for each image patch. We first evaluate locality alignment with a vision-only benchmark, finding that it improves a model’s performance at patch-level semantic segmentation, especially for strong backbones trained with image-caption pairs (e.g., CLIP and SigLIP). We then train a series of VLMs with and without locality alignment, and show that locality-aligned backbones improve performance across a range of benchmarks, particularly ones that involve spatial understanding (e.g., RefCOCO, OCID-Ref, TallyQA, VSR, AI2D). Overall, we demonstrate that we can efficiently learn local semantic extraction via a locality alignment stage, and that this procedure benefits VLM training recipes that use off-the-shelf vision backbones.

# 2409. Robust LLM safeguarding via refusal feature adversarial training

链接：https://iclr.cc/virtual/2025/poster/28144 abstract： Large language models (LLMs) are vulnerable to adversarial attacks that can elicit harmful responses. Defending against such attacks remains challenging due to the opacity of jailbreaking mechanisms and the high computational cost of training LLMs robustly. We demonstrate that adversarial attacks share a universal mechanism for circumventing LLM safeguards that works by ablating a dimension in the residual stream embedding

space called the refusal feature. We further show that the operation of refusal feature ablation (RFA) approximates the worst-case perturbation of offsetting model safety. Based on these findings, we propose Refusal Feature Adversarial Training (ReFAT), a novel algorithm that efficiently performs LLM adversarial training by simulating the effect of input-level attacks via RFA. Experiment results show that ReFAT significantly improves the robustness of three popular LLMs against a wide range of adversarial attacks, with considerably less computational overhead compared to existing adversarial training methods.

## 2410. UniCO: On Unified Combinatorial Optimization via Problem Reduction to Matrix-Encoded General TSP

链接：https://iclr.cc/virtual/2025/poster/27742 abstract： Various neural solvers have been devised for combinatorial optimization (CO), which are often tailored for specific problem types, e.g., TSP, CVRP and SAT, etc. Yet, it remains an open question how to achieve universality regarding problem representing and learning with a general framework. This paper first proposes UniCO, to unify a set of CO problems by reducing them into the general TSP form featured by distance matrices. The applicability of this strategy depends on the efficiency of the problem reduction and solution transition procedures, which we show that at least ATSP, HCP, and SAT are readily feasible. The hope is to allow for the effective and even simultaneous use of as many types of CO instances as possible to train a neural TSP solver, and optionally finetune it for specific problem types. In particular, unlike the prevalent TSP benchmarks based on Euclidean instances with 2-D coordinates, our studied domain of TSP could involve non-metric, asymmetric or discrete distances without explicit node coordinates, which is much less explored in TSP literature while poses new intellectual challenges. Along this direction, we devise two neural TSP solvers with and without supervision to conquer such matrix-formulated input, respectively: 1) MatPOENet and 2) MatDIFFNet. The former is a reinforcement learning-based sequential model with pseudo one-hot embedding (POE) scheme; and the latter is a Diffusion-based generative model with the mix-noised reference mapping scheme. Experiments on ATSP, 2DTSP, HCP- and SAT-distributed general TSPs show the strong ability towards arbitrary matrix-encoded TSP with structure and size variation.

## 2411. Lasso Bandit with Compatibility Condition on Optimal Arm

链接：https://iclr.cc/virtual/2025/poster/28898 abstract： We consider a stochastic sparse linear bandit problem where only a sparse subset of context features affects the expected reward function, i.e., the unknown reward parameter has a sparse structure.In the existing Lasso bandit literature, the compatibility conditions, together with additional diversity conditions on the context features are imposed to achieve regret bounds that only depend logarithmically on the ambient dimension $d$.In this paper, we demonstrate that even without the additional diversity assumptions, the \textit{compatibility condition on the optimal arm} is sufficient to derive a regret bound that depends logarithmically on $d$, and our assumption is strictly weaker than those used in the lasso bandit literature under the single-parameter setting.We propose an algorithm that adapts the forced-sampling technique and prove that the proposed algorithm achieves $\mathcal{O}(\text{poly}\log dT)$ regret under the margin condition.To our knowledge, the proposed algorithm requires the weakest assumptions among Lasso bandit algorithms under the single-parameter setting that achieve $\mathcal{O}(\text{poly}\log dT)$ regret.Through numerical experiments, we confirm the superior performance of our proposed algorithm.

## 2412. Preference Elicitation for Offline Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/31114 abstract： Applying reinforcement learning (RL) to real-world problems is often made challenging by the inability to interact with the environment and the difficulty of designing reward functions. Offline RL addresses the first challenge by considering access to an offline dataset of environment interactions labeled by the reward function. In contrast, Preference-based RL does not assume access to the reward function and learns it from preferences, but typically requires an online interaction with the environment. We bridge the gap between these frameworks by exploring efficient methods for acquiring preference feedback in a fully offline setup. We propose Sim-OPRL, an offline preference-based reinforcement learning algorithm, which leverages a learned environment model to elicit preference feedback on simulated rollouts. Drawing on insights from both the offline RL and the preference-based RL literature, our algorithm employs a pessimistic approach for out-of-distribution data, and an optimistic approach for acquiring informative preferences about the optimal policy. We provide theoretical guarantees regarding the sample complexity of our approach, dependent on how well the offline data covers the optimal policy. Finally, we demonstrate the empirical performance of Sim-OPRL in various environments.

## 2413. Unify ML4TSP: Drawing Methodological Principles for TSP and Beyond from Streamlined Design Space of Learning and Search

链接：https://iclr.cc/virtual/2025/poster/28796 abstract： Despite the rich works on machine learning (ML) for combinatorial optimization (CO), a unified, principled framework remains lacking. This study utilizes the Travelling Salesman Problem (TSP) as a major case study, with adaptations demonstrated for other CO problems, dissecting established mainstream learning-based solvers to outline a comprehensive design space. We present ML4TSPBench, which advances a unified modular streamline incorporating existing technologies in both learning and search for transparent ablation, aiming to reassess the role of learning and discern which parts of existing techniques are genuinely beneficial and which are not. This further leads to the investigation of desirable principles of learning designs and the exploration of concepts guiding method designs. We demonstrate the desirability of principles such as joint probability estimation, symmetry solution representation, and online optimization for learning-based designs. Leveraging the findings, we propose enhancements to existing methods to compensate for their missing attributes, thereby advancing performance and enriching the technique library. From a higher

viewpoint, we also uncover a performance advantage in non-autoregressive and supervised paradigms compared to their counterparts. The strategic decoupling and organic recompositions yield a factory of new TSP solvers, where we investigate synergies across various method combinations and pinpoint the optimal design choices to create more powerful ML4TSP solvers, thereby facilitating and offering a reference for future research and engineering endeavors.

## 2414. Guaranteed Generation from Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30739 abstract： As large language models (LLMs) are increasingly used across various applications, there is a growing need to control text generation to satisfy specific constraints or requirements. This raises a crucial question: Is it possible to guarantee strict constraint satisfaction in generated outputs while preserving the distribution of the original model as much as possible? We first define the ideal distribution — the one closest to the original model, which also always satisfies the expressed constraint — as the ultimate goal of guaranteed generation. We then state a fundamental limitation, namely that it is impossible to reach that goal through autoregressive training alone. This motivates the necessity of combining training-time and inference-time methods to enforce such guarantees. Based on this insight, we propose GUARD, a simple yet effective approach that combines an autoregressive proposal distribution with rejection sampling. Through GUARD's theoretical properties, we show how controlling the KL divergence between a specific proposal and the target ideal distribution simultaneously optimizes inference speed and distributional closeness. To validate these theoretical concepts, we conduct extensive experiments on two text generation settings with hard-to-satisfy constraints: a lexical constraint scenario and a sentiment reversal scenario. These experiments show that GUARD achieves perfect constraint satisfaction while almost preserving the ideal distribution with highly improved inference efficiency. GUARD provides a principled approach to enforcing strict guarantees for LLMs without compromising their generative capabilities.

## 2415. Forget the Data and Fine-Tuning! Just Fold the Network to Compress

链接：https://iclr.cc/virtual/2025/poster/29387 abstract： We introduce model folding, a novel data-free model compression technique that merges structurally similar neurons across layers, significantly reducing the model size without the need for fine-tuning or access to training data. Unlike existing methods, model folding preserves data statistics during compression by leveraging k-means clustering, and using novel data-free techniques to prevent variance collapse or explosion. Our theoretical framework and experiments across standard benchmarks, including ResNet18 and LLaMA-7B, demonstrate that model folding achieves comparable performance to data-driven compression techniques and outperforms recently proposed data-free methods, especially at high sparsity levels. This approach is particularly effective for compressing large-scale models, making it suitable for deployment in resource-constrained environments.

## 2416. Statistical Advantages of Perturbing Cosine Router in Mixture of Experts

链接：https://iclr.cc/virtual/2025/poster/28871 abstract： The cosine router in Mixture of Experts (MoE) has recently emerged as an attractive alternative to the conventional linear router. Indeed, the cosine router demonstrates favorable performance in image and language tasks and exhibits better ability to mitigate the representation collapse issue, which often leads to parameter redundancy and limited representation potentials. Despite its empirical success, a comprehensive analysis of the cosine router in MoE has been lacking. Considering the least square estimation of the cosine routing MoE, we demonstrate that due to the intrinsic interaction of the model parameters in the cosine router via some partial differential equations, regardless of the structures of the experts, the estimation rates of experts and model parameters can be as slow as $\mathcal{O}(1/\log^{\tau}(n))$ where $\tau > 0$ is some constant and $n$ is the sample size. Surprisingly, these pessimistic non-polynomial convergence rates can be circumvented by the widely used technique in practice to stabilize the cosine router --- simply adding noises to the $\ell^2$-norms in the cosine router, which we refer to as *perturbed cosine router*. Under the strongly identifiable settings of the expert functions, we prove that the estimation rates for both the experts and model parameters under the perturbed cosine routing MoE are significantly improved to polynomial rates. Finally, we conduct extensive simulation studies in both synthetic and real data settings to empirically validate our theoretical results.

## 2417. Zero-Shot Natural Language Explanations

链接：https://iclr.cc/virtual/2025/poster/29324 abstract： Natural Language Explanations (NLEs) interpret the decision-making process of a given model through textual sentences. Current NLEs suffer from a severe limitation; they are unfaithful to the model's actual reasoning process, as a separate textual decoder is explicitly trained to generate those explanations using annotated datasets for a specific task, leading them to reflect what annotators desire. In this work, we take the first step towards generating faithful NLEs for any visual classification model without any training data. Our approach models the relationship between class embeddings from the classifier of the vision model and their corresponding class names via a simple MLP which trains in seconds. After training, we can map any new text to the classifier space and measure its association with the visual features. We conduct experiments on 38 vision models, including both CNNs and Transformers. In addition to NLEs, our method offers other advantages such as zero-shot image classification and fine-grained concept discovery.

## 2418. Learning Geometric Reasoning Networks For Robot Task And Motion Planning

链接：https://iclr.cc/virtual/2025/poster/29152 abstract： Task and Motion Planning (TAMP) is a computationally challenging robotics problem due to the tight coupling of discrete symbolic planning and continuous geometric planning of robot motions. In particular, planning manipulation tasks in complex 3D environments leads to a large number of costly geometric planner queries to verify the feasibility of considered actions and plan their motions. To address this issue, we propose Geometric Reasoning Networks (GRN), a graph neural network (GNN)-based model for action and grasp feasibility prediction, designed to significantly reduce the dependency on the geometric planner. Moreover, we introduce two key interpretability mechanisms: inverse kinematics (IK) feasibility prediction and grasp obstruction (GO) estimation. These modules not only improve feasibility predictions accuracy, but also explain why certain actions or grasps are infeasible, thus allowing a more efficient search for a feasible solution. Through extensive experimental results, we show that our model outperforms state-of-the-art methods, while maintaining generalizability to more complex environments, diverse object shapes, multi-robot settings, and real-world robots.

## 2419. Learning to Help in Multi-Class Settings

链接：https://iclr.cc/virtual/2025/poster/29894 abstract： Deploying complex machine learning models on resource-constrained devices is challenging due to limited computational power, memory, and model retrainability. To address these limitations, a hybrid system can be established by augmenting the local model with a server-side model, where samples are selectively deferred by a rejector and then sent to the server for processing. The hybrid system enables efficient use of computational resources while minimizing the overhead associated with server usage. The recently proposed Learning to Help (L2H) model proposed training a server model given a fixed local (client) model. This differs from the Learning to Defer (L2D) framework which trains the client for a fixed (expert) server. In both L2D and L2H, the training includes learning a rejector at the client to determine when to query the server. In this work, we extend the L2H model from binary to multi-class classification problems and demonstrate its applicability in a number of different scenarios of practical interest in which access to the server may be limited by cost, availability, or policy. We derive a stage-switching surrogate loss function that is differentiable, convex, and consistent with the Bayes rule corresponding to the 0-1 loss for the L2H model. Experiments show that our proposed methods offer an efficient and practical solution for multi-class classification in resource-constrained environments.

## 2420. Variational Diffusion Posterior Sampling with Midpoint Guidance

链接：https://iclr.cc/virtual/2025/poster/30902 abstract： Diffusion models have recently shown considerable potential in solving Bayesian inverse problems when used as priors. However, sampling from the resulting denoising posterior distributions remains a challenge as it involves intractable terms. To tackle this issue, state-of-the-art approaches formulate the problem as that of sampling from a surrogate diffusion model targeting the posterior and decompose its scores into two terms: the prior score and an intractable guidance term. While the former is replaced by the pre-trained score of the considered diffusion model, the guidance term has to be estimated. In this paper, we propose a novel approach that utilises a decomposition of the transitions which, in contrast to previous methods, allows a trade-off between the complexity of the intractable guidance term and that of the prior transitions. We validate the proposed approach through extensive experiments on linear and nonlinear inverse problems, including challenging cases with latent diffusion models as priors, and demonstrate its effectiveness in reconstructing electrocardiogram (ECG) from partial measurements for accurate cardiac diagnosis.

## 2421. Disentangling Representations through Multi-task Learning

链接：https://iclr.cc/virtual/2025/poster/27723 abstract： Intelligent perception and interaction with the world hinges on internal representations that capture its underlying structure ("disentangled" or "abstract" representations). Disentangled representations serve as world models, isolating latent factors of variation in the world along approximately orthogonal directions, thus facilitating feature-based generalization. We provide experimental and theoretical results guaranteeing the emergence of disentangled representations in agents that optimally solve multi-task evidence accumulation classification tasks, canonical in the neuroscience literature. The key conceptual finding is that, by producing accurate multi-task classification estimates, a system implicitly represents a set of coordinates specifying a disentangled representation of the underlying latent state of the data it receives. The theory provides conditions for the emergence of these representations in terms of noise, number of tasks, and evidence accumulation time, when the classification boundaries are affine in the latent space. Surprisingly, the theory also produces closed-form expressions for extracting the disentangled representation from the model's latent state $\mathbf Z(t)$. We experimentally validate these predictions in RNNs trained on multi-task classification, which learn disentangled representations in the form of continuous attractors, leading to zero-shot out-of-distribution (OOD) generalization in predicting latent factors. We demonstrate the robustness of our framework across autoregressive architectures, decision boundary geometries and in tasks requiring classification confidence estimation. We find that transformers are particularly suited for disentangling representations, which might explain their unique world understanding abilities. Overall, our framework establishes a formal link between competence at multiple tasks and the formation of disentangled, interpretable world models in both biological and artificial systems, and helps explain why ANNs often arrive at human-interpretable concepts, and how they both may acquire exceptional zero-shot generalization capabilities.

## 2422. Prompting Fairness: Integrating Causality to Debias Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30833 abstract： Large language models (LLMs), despite their remarkable capabilities, are susceptible to generating biased and discriminatory responses. As LLMs increasingly influence high-stakes decision-making (e.g., hiring and healthcare), mitigating these biases becomes critical. In this work, we propose a causality-guided

debiasing framework to tackle social biases, aiming to reduce the objectionable dependence between LLMs' decisions and the social information in the input. Our framework introduces a novel perspective to identify how social information can affect an LLM's decision through different causal pathways. Leveraging these causal insights, we outline principled prompting strategies that regulate these pathways through selection mechanisms. This framework not only unifies existing prompting-based debiasing techniques, but also opens up new directions for reducing bias by encouraging the model to prioritize fact-based reasoning over reliance on biased social cues. We validate our framework through extensive experiments on real-world datasets across multiple domains, demonstrating its effectiveness in debiasing LLM decisions, even with only black-box access to the model.

## 2423. T-JEPA: Augmentation-Free Self-Supervised Learning for Tabular Data

链接：https://iclr.cc/virtual/2025/poster/28793 abstract： Self-supervision is often used for pre-training to foster performance on a downstream task by constructing meaningful representations of samples. Self-supervised learning (SSL) generally involves generating different views of the same sample and thus requires data augmentations that are challenging to construct for tabular data. This constitutes one of the main challenges of self-supervision for structured data. In the present work, we propose a novel augmentation-free SSL method for tabular data. Our approach, T-JEPA, relies on a Joint Embedding Predictive Architecture (JEPA) and is akin to mask reconstruction in the latent space. It involves predicting the latent representation of one subset of features from the latent representation of a different subset within the same sample, thereby learning rich representations without augmentations. We use our method as a pre-training technique and train several deep classifiers on the obtained representation. Our experimental results demonstrate a substantial improvement in both classification and regression tasks, outperforming models trained directly on samples in their original data space. Moreover, T-JEPA enables some methods to consistently outperform or match the performance of traditional methods likes Gradient Boosted Decision Trees. To understand why, we extensively characterize the obtained representations and show that T-JEPA effectively identifies relevant features for downstream tasks without access to the labels. Additionally, we introduce regularization tokens, a novel regularization method critical for training of JEPA-based models on structured data.

## 2424. Causal Discovery via Bayesian Optimization

链接：https://iclr.cc/virtual/2025/poster/30746 abstract： Existing score-based methods for directed acyclic graph (DAG) learning from observational data struggle to recover the causal graph accurately and sample-efficiently. To overcome this, in this study, we propose DrBO (DAG recovery via Bayesian Optimization)—a novel DAG learning framework leveraging Bayesian optimization (BO) to find high-scoring DAGs. We show that, by sophisticatedly choosing the promising DAGs to explore, we can find higher-scoring ones much more efficiently. To address the scalability issues of conventional BO in DAG learning, we replace Gaussian Processes commonly employed in BO with dropout neural networks, trained in a continual manner, which allows for (i) flexibly modeling the DAG scores without overfitting, (ii) incorporation of uncertainty into the estimated scores, and (iii) scaling with the number of evaluations. As a result, DrBO is computationally efficient and can find the accurate DAG in fewer trials and less time than existing state-of-the-art methods. This is demonstrated through an extensive set of empirical evaluations on many challenging settings with both synthetic and real data. Our implementation is available at https://github.com/baosws/DrBO.

## 2425. Deep Incomplete Multi-view Learning via Cyclic Permutation of VAEs

链接：https://iclr.cc/virtual/2025/poster/28147 abstract： Multi-View Representation Learning (MVRL) aims to derive a unified representation from multi-view data by leveraging shared and complementary information across views. However, when views are irregularly missing, the incomplete data can lead to representations that lack sufficiency and consistency. To address this, we propose Multi-View Permutation of Variational Auto-Encoders (MVP), which excavates invariant relationships between views in incomplete data. MVP establishes inter-view correspondences in the latent space of Variational Auto-Encoders, enabling the inference of missing views and the aggregation of more sufficient information. To derive a valid Evidence Lower Bound (ELBO) for learning, we apply permutations to randomly reorder variables for cross-view generation and then partition them by views to maintain invariant meanings under permutations. Additionally, we enhance consistency by introducing an informational prior with cyclic permutations of posteriors, which turns the regularization term into a similarity measure across distributions. We demonstrate the effectiveness of our approach on seven diverse datasets with varying missing ratios, achieving superior performance in multi-view clustering and generation tasks.

## 2426. Contrastive Learning from Synthetic Audio Doppelgängers

链接：https://iclr.cc/virtual/2025/poster/29305 abstract： Learning robust audio representations currently demands extensive datasets of real-world sound recordings. By applying artificial transformations to these recordings, models can learn to recognize similarities despite subtle variations through techniques like contrastive learning. However, these transformations are only approximations of the true diversity found in real-world sounds, which are generated by complex interactions of physical processes, from vocal cord vibrations to the resonance of musical instruments. We propose a solution to both the data scale and transformation limitations, leveraging synthetic audio. By randomly perturbing the parameters of a sound synthesizer, we generate audio doppelgängers—synthetic positive pairs with causally manipulated variations in timbre, pitch, and temporal envelopes. These variations, difficult to achieve through augmentations of existing audio, provide a rich source of contrastive information. Despite the shift to randomly generated synthetic data, our method produces strong representations, outperforming real data on several standard audio classification tasks. Notably, our approach is lightweight, requires no data storage, and has only a single hyperparameter, which we extensively analyze. We offer this method as a complement to existing strategies for

contrastive learning in audio, using synthesized sounds to reduce the data burden on practitioners.

## 2427. When Selection Meets Intervention: Additional Complexities in Causal Discovery

链接：https://iclr.cc/virtual/2025/poster/27799 abstract： We address the common yet often-overlooked selection bias in interventional studies, where subjects are selectively enrolled into experiments. For instance, participants in a drug trial are usually patients of the relevant disease; A/B tests on mobile applications target existing users only, and gene perturbation studies typically focus on specific cell types, such as cancer cells. Ignoring this bias leads to incorrect causal discovery results. Even when recognized, the existing paradigm for interventional causal discovery still fails to address it. This is because subtle differences in when and where interventions happen can lead to significantly different statistical patterns. We capture this dynamic by introducing a graphical model that explicitly accounts for both the observed world (where interventions are applied) and the counterfactual world (where selection occurs while interventions have not been applied). We characterize the Markov property of the model, and propose a provably sound algorithm to identify causal relations as well as selection mechanisms up to the equivalence class, from data with soft interventions and unknown targets. Through synthetic and real-world experiments, we demonstrate that our algorithm effectively identifies true causal relations despite the presence of selection bias.

## 2428. AVHBench: A Cross-Modal Hallucination Benchmark for Audio-Visual Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28638 abstract： Following the success of Large Language Models (LLMs), expanding their boundaries to new modalities represents a significant paradigm shift in multimodal understanding. Human perception is inherently multimodal, relying not only on text but also on auditory and visual cues for a complete understanding of the world. In recognition of this fact, audio-visual LLMs have recently emerged. Despite promising developments, the lack of dedicated benchmarks poses challenges for understanding and evaluating models. In this work, we show that audio-visual LLMs struggle to discern subtle relationships between audio and visual signals, leading to hallucinations and highlighting the need for reliable benchmarks. To address this, we introduce AVHBench, the first comprehensive benchmark specifically designed to evaluate the perception and comprehension capabilities of audio-visual LLMs. Our benchmark includes tests for assessing hallucinations, as well as the cross-modal matching and reasoning abilities of these models. Our results reveal that most existing audio-visual LLMs struggle with hallucinations caused by cross-interactions between modalities, due to their limited capacity to perceive complex multimodal signals and their relationships. Additionally, we demonstrate that simple training with our AVHBench improves robustness of audio-visual LLMs against hallucinations. Dataset: https://github.com/kaist-ami/AVHBench

## 2429. Dynamic Negative Guidance of Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/30862 abstract： Negative Prompting (NP) is widely utilized in diffusion models, particularly in text-to-image applications, to prevent the generation of undesired features. In this paper, we show that conventional NP is limited by the assumption of a constant guidance scale, which may lead to highly suboptimal results, or even complete failure, due to the non-stationarity and state-dependence of the reverse process. Based on this analysis, we derive a principled technique called Dynamic Negative Guidance, which relies on a near-optimal time and state dependent modulation of the guidance without requiring additional training. Unlike NP, negative guidance requires estimating the posterior class probability during the denoising process, which is achieved with limited additional computational overhead by tracking the discrete Markov Chain during the generative process. We evaluate the performance of DNG class-removal on MNIST and CIFAR10, where we show that DNG leads to higher safety, preservation of class balance and image quality when compared with baseline methods. Furthermore, we show that it is possible to use DNG with Stable Diffusion to obtain more accurate and less invasive guidance than NP.

## 2430. Bilinear MLPs enable weight-based mechanistic interpretability

链接：https://iclr.cc/virtual/2025/poster/28827 abstract： A mechanistic understanding of how MLPs do computation in deep neural net-works remains elusive. Current interpretability work can extract features fromhidden activations over an input dataset but generally cannot explain how MLPweights construct features. One challenge is that element-wise nonlinearitiesintroduce higher-order interactions and make it difficult to trace computationsthrough the MLP layer. In this paper, we analyze bilinear MLPs, a type ofGated Linear Unit (GLU) without any element-wise nonlinearity that neverthe-less achieves competitive performance. Bilinear MLPs can be fully expressed interms of linear operations using a third-order tensor, allowing flexible analysis ofthe weights. Analyzing the spectra of bilinear MLP weights using eigendecom-position reveals interpretable low-rank structure across toy tasks, image classifi-cation, and language modeling. We use this understanding to craft adversarialexamples, uncover overfitting, and identify small language model circuits directlyfrom the weights alone. Our results demonstrate that bilinear layers serve as aninterpretable drop-in replacement for current activation functions and that weight-based interpretability is viable for understanding deep-learning models.

## 2431. Super(ficial)-alignment: Strong Models May Deceive Weak Models in Weak-to-Strong Generalization

链接：https://iclr.cc/virtual/2025/poster/30200 abstract： Superalignment, where humans act as weak supervisors for superhuman models, has become a crucial problem with the rapid development of Large Language Models (LLMs). Recent work has preliminarily studied this problem by using weak models to supervise strong models, and discovered that weakly supervised strong students can consistently outperform weak teachers towards the alignment target, leading to a weak-to-strong generalization phenomenon. However, we are concerned that behind such a promising phenomenon, whether there exists an issue of weak-to-strong deception, where strong models deceive weak models by exhibiting well-aligned in areas known to weak models but producing misaligned behaviors in cases weak models do not know. We take an initial step towards exploring this security issue in a specific but realistic multi-objective alignment case, where there may be some alignment targets conflicting with each other (e.g., helpfulness v.s. harmlessness). We aim to explore whether, in such cases, strong models might deliberately make mistakes in areas known to them but unknown to weak models within one alignment dimension, in exchange for a higher reward in another dimension. Through extensive experiments in both the reward modeling and preference optimization scenarios, we find: (1) The weak-to-strong deception phenomenon exists across all settings. (2) The deception intensifies as the capability gap between weak and strong models increases. (3) Bootstrapping with an intermediate model can mitigate the deception to some extent, though its effectiveness remains limited. Our work highlights the urgent need to pay more attention to the true reliability of superalignment.

## 2432. Direct Post-Training Preference Alignment for Multi-Agent Motion Generation Model Using Implicit Feedback from Pre-training Demonstrations

链接：https://iclr.cc/virtual/2025/poster/30765 abstract： Recent advancements in Large Language Models (LLMs) have revolutionized motion generation models in embodied applications such as autonomous driving and robotic manipulation. While LLM-type auto-regressive motion generation models benefit from training scalability, there remains a discrepancy between their token prediction objectives and human preferences. As a result, models pre-trained solely with token-prediction objectives often generate behaviors that deviate from what humans would prefer, making post-training preference alignment crucial for producing human-preferred motions. Unfortunately, post-training alignment requires extensive preference rankings of motions generated by the pre-trained model, which are costly and time-consuming to annotate, especially in multi-agent motion generation settings. Recently, there has been growing interest in leveraging expert demonstrations previously used during pre-training to scalably generate preference data for post-training alignment. However, these methods often adopt an adversarial assumption, treating all pre-trained model-generated samples as unpreferred examples and relying solely on pre-training expert demonstrations to construct preferred examples. This adversarial approach overlooks the valuable signal provided by preference rankings among the model's own generations, ultimately reducing alignment effectiveness and potentially leading to misaligned behaviors. In this work, instead of treating all generated samples as equally bad, we propose a principled approach that leverages implicit preferences encoded in pre-training expert demonstrations to construct preference rankings among the pre-trained model's generations, offering more nuanced preference alignment guidance with zero human cost. We apply our approach to large-scale traffic simulation (more than 100 agents) and demonstrate its effectiveness in improving the realism of pre-trained model's generated behaviors, making a lightweight 1M motion generation model comparable to state-of-the-art large imitation-based models by relying solely on implicit feedback from pre-training demonstrations, without requiring additional post-training human preference annotations or incurring high computational costs. Furthermore, we provide an in-depth analysis of preference data scaling laws and their effects on over-optimization, offering valuable insights for future studies.

## 2433. SaRA: High-Efficient Diffusion Model Fine-tuning with Progressive Sparse Low-Rank Adaptation

链接：https://iclr.cc/virtual/2025/poster/27860 abstract： The development of diffusion models has led to significant progress in image and video generation tasks, with pre-trained models like the Stable Diffusion series playing a crucial role.However, a key challenge remains in downstream task applications: how to effectively and efficiently adapt pre-trained diffusion models to new tasks.Inspired by model pruning which lightens large pre-trained models by removing unimportant parameters, we propose a novel model fine-tuning method to make full use of these ineffective parameters and enable the pre-trained model with new task-specified capabilities.In this work, we first investigate the importance of parameters in pre-trained diffusion models and discover that parameters with the smallest absolute values do not contribute to the generation process due to training instabilities.Based on this observation, we propose a fine-tuning method termed SaRA that re-utilizes these temporarily ineffective parameters, equating to optimizing a sparse weight matrix to learn the task-specific knowledge.To mitigate potential overfitting, we propose a nuclear-norm-based low-rank sparse training scheme for efficient fine-tuning.Furthermore, we design a new progressive parameter adjustment strategy to make full use of the finetuned parameters.Finally, we propose a novel unstructural backpropagation strategy, which significantly reduces memory costs during fine-tuning.Our method enhances the generative capabilities of pre-trained models in downstream applications and outperforms existing fine-tuning methods in maintaining model's generalization ability. Source code is available at https://sjtuplayer.github.io/projects/SaRA.

## 2434. MediConfusion: Can you trust your AI radiologist? Probing the reliability of multimodal medical foundation models

链接：https://iclr.cc/virtual/2025/poster/30242 abstract： Multimodal Large Language Models (MLLMs) have tremendous potential to improve the accuracy, availability, and cost-effectiveness of healthcare by providing automated solutions or serving as aids to medical professionals. Despite promising first steps in developing medical MLLMs in the past few years, their capabilities and limitations are not well understood. Recently, many benchmark datasets have been proposed that test the

general medical knowledge of such models across a variety of medical areas. However, the systematic failure modes and vulnerabilities of such models are severely underexplored with most medical benchmarks failing to expose the shortcomings of existing models in this safety-critical domain. In this paper, we introduce MediConfusion, a challenging medical Visual Question Answering (VQA) benchmark dataset, that probes the failure modes of medical MLLMs from a vision perspective. We reveal that state-of-the-art models are easily confused by image pairs that are otherwise visually dissimilar and clearly distinct for medical experts. Strikingly, all available models (open-source or proprietary) achieve performance below random guessing on MediConfusion, raising serious concerns about the reliability of existing medical MLLMs for healthcare deployment. We also extract common patterns of model failure that may help the design of a new generation of more trustworthy and reliable MLLMs in healthcare.

## 2435. AdaFisher: Adaptive Second Order Optimization via Fisher Information

链接：https://iclr.cc/virtual/2025/poster/28275 abstract： First-order optimization methods are currently the mainstream in training deep neural networks (DNNs). Optimizers like Adam incorporate limited curvature information by employing the diagonal matrix preconditioning of the stochastic gradient during the training. Despite their widespread, second-order optimization algorithms exhibit superior convergence properties compared to their first-order counterparts e.g. Adam and SGD. However, their practicality in training DNNs is still limited due to increased per-iteration computations compared to the first-order methods. We present AdaFisher--an adaptive second-order optimizer that leverages a diagonal block-Kronecker approximation of the Fisher information matrix for adaptive gradient preconditioning. AdaFisher aims to bridge the gap between enhanced convergence/generalization capabilities and computational efficiency in second-order optimization framework for training DNNs. Despite the slow pace of second-order optimizers, we showcase that AdaFisher can be reliably adopted for image classification, language modeling and stands out for its stability and robustness in hyper-parameter tuning. We demonstrate that AdaFisher outperforms the SOTA optimizers in terms of both accuracy and convergence speed. Code is available from https://github.com/AtlasAnalyticsLab/AdaFisher.

## 2436. DocMIA: Document-Level Membership Inference Attacks against DocVQA Models

链接：https://iclr.cc/virtual/2025/poster/28823 abstract： Document Visual Question Answering (DocVQA) has introduced a new paradigm for end-to-end document understanding, and quickly became one of the standard benchmarks for multimodal LLMs. Automating document processing workflows, driven by DocVQA models, presents significant potential for many business sectors. However, documents tend to contain highly sensitive information, raising concerns about privacy risks associated with training such DocVQA models. One significant privacy vulnerability, exploited by the membership inference attack, is the possibility for an adversary to determine if a particular record was part of the model's training data. In this paper, we introduce two novel membership inference attacks tailored specifically to DocVQA models. These attacks are designed for two different adversarial scenarios: a white-box setting, where the attacker has full access to the model architecture and parameters, and a black-box setting, where only the model's outputs are available. Notably, our attacks assume the adversary lacks access to auxiliary datasets, which is more realistic in practice but also more challenging. Our unsupervised methods outperform existing state-of-the-art membership inference attacks across a variety of DocVQA models and datasets, demonstrating their effectiveness and highlighting the privacy risks in this domain.

## 2437. The Directionality of Optimization Trajectories in Neural Networks

链接：https://iclr.cc/virtual/2025/poster/30103 abstract： The regularity or implicit bias in neural network optimization has been typically studied via the parameter norms or the landscape curvature, often overlooking the trajectory leading to these parameters. However, properties of the trajectory --- particularly its directionality --- capture critical aspects of how gradient descent navigates the landscape to converge to a solution. In this work, we introduce the notion of a Trajectory Map and derive natural complexity measures that highlight the directional characteristics of optimization trajectories. Our comprehensive analysis across vision and language modeling tasks reveals that (a) the trajectory's directionality at the macro-level saturates by the initial phase of training, wherein weight decay and momentum play a crucial but understated role; and (b) in subsequent training, trajectory directionality manifests in micro-level behaviors, such as oscillations, for which we also provide a theoretical analysis. This implies that neural optimization trajectories have, overall, a more linear form than zig-zaggy, as evident by high directional similarity, especially towards the end. To further hone this point, we show that when the trajectory direction gathers such an inertia, optimization proceeds largely unaltered even if the network is severely decapacitated (by freezing >99% of the parameters), --- thereby demonstrating the potential for significant computational and resource savings without compromising performance.

## 2438. Expressivity of Neural Networks with Random Weights and Learned Biases

链接：https://iclr.cc/virtual/2025/poster/30920 abstract： Landmark universal function approximation results for neural networks with trained weights and biases provided the impetus for the ubiquitous use of neural networks as learning models in neuroscience and Artificial Intelligence (AI). Recent work has extended these results to networks in which a smaller subset of weights (e.g., output weights) are tuned, leaving other parameters random. However, it remains an open question whether universal approximation holds when only biases are learned, despite evidence from neuroscience and AI that biases

significantly shape neural responses. The current paper answers this question. We provide theoretical and numerical evidence demonstrating that feedforward neural networks with fixed random weights can approximate any continuous function on compact sets. We further show an analogous result for the approximation of dynamical systems with recurrent neural networks. Our findings are relevant to neuroscience, where they demonstrate the potential for behaviourally relevant changes in dynamics without modifying synaptic weights, as well as for AI, where they shed light on recent fine-tuning methods for large language models, like bias and prefix-based approaches.

## 2439. Beyond Random Augmentations: Pretraining with Hard Views

链接：https://iclr.cc/virtual/2025/poster/30639 abstract： Self-Supervised Learning (SSL) methods typically rely on random image augmentations, or views, to make models invariant to different transformations. We hypothesize that the efficacy of pretraining pipelines based on conventional random view sampling can be enhanced by explicitly selecting views that benefit the learning progress. A simple yet effective approach is to select hard views that yield a higher loss. In this paper, we propose Hard View Pretraining (HVP), a learning-free strategy that extends random view generation by exposing models to more challenging samples during SSL pretraining. HVP encompasses the following iterative steps: 1) randomly sample multiple views and forward each view through the pretrained model, 2) create pairs of two views and compute their loss, 3) adversarially select the pair yielding the highest loss according to the current model state, and 4) perform a backward pass with the selected pair. In contrast to existing hard view literature, we are the first to demonstrate hard view pretraining's effectiveness at scale, particularly training on the full ImageNet-1k dataset, and evaluating across multiple SSL methods, Convolutional Networks, and Vision Transformers. As a result, HVP sets a new state-of-the-art on DINO ViT-B/16, reaching 78.8% linear evaluation accuracy (a 0.6% improvement) and consistent gains of 1% for both 100 and 300 epoch pretraining, with similar improvements across transfer tasks in DINO, SimSiam, iBOT, and SimCLR.

## 2440. Attention layers provably solve single-location regression

链接：https://iclr.cc/virtual/2025/poster/30454 abstract： Attention-based models, such as Transformer, excel across various tasks but lack a comprehensive theoretical understanding, especially regarding token-wise sparsity and internal linear representations. To address this gap, we introduce the single-location regression task, where only one token in a sequence determines the output, and its position is a latent random variable, retrievable via a linear projection of the input. To solve this task, we propose a dedicated predictor, which turns out to be a simplified version of a non-linear self-attention layer. We study its theoretical properties, by showing its asymptotic Bayes optimality and analyzing its training dynamics. In particular, despite the non-convex nature of the problem, the predictor effectively learns the underlying structure. This work highlights the capacity of attention mechanisms to handle sparse token information and internal linear structures.

## 2441. Fine-Tuning Token-Based Large Multimodal Models: What Works, What Doesn’t and What's Next

链接：https://iclr.cc/virtual/2025/poster/31328 abstract： In this blog post, we explore the advancements and challenges in fine-tuning unified token-based large multimodal models, focusing on the Chameleon architecture and its fine-tuned variant, Anole. Released in 2024, these models exemplify a modern approach for integrating various data modalities through tokens, simplifying modal fusion and leveraging established techniques from large language models. The post details our research efforts to reveal what is important, what is mistaken, and what is worth exploring in future research during the fine-tuning process.

## 2442. Lines of Thought in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/27654 abstract： Large Language Models achieve next-token prediction by transporting a vectorized piece of text (prompt) across an accompanying embedding space under the action of successive transformer layers. The resulting high-dimensional trajectories realize different contextualization, or 'thinking', steps, and fully determine the output probability distribution. We aim to characterize the statistical properties of ensembles of these 'lines of thought.' We observe that independent trajectories cluster along a low-dimensional, non-Euclidean manifold, and that their path can be well approximated by a stochastic equation with few parameters extracted from data. We find it remarkable that the vast complexity of such large models can be reduced to a much simpler form, and we reflect on implications.

## 2443. Density estimation with LLMs: a geometric investigation of in-context learning trajectories

链接：https://iclr.cc/virtual/2025/poster/28109 abstract： Large language models (LLMs) demonstrate remarkable emergent abilities to perform in-context learning across various tasks, including time series forecasting. This work investigates LLMs' ability to estimate probability density functions (PDFs) from data observed in-context; such density estimation (DE) is a fundamental task underlying many probabilistic modeling problems. We leverage the Intensive Principal Component Analysis (InPCA) to visualize and analyze the in-context learning dynamics of LLaMA-2 models. Our main finding is that these LLMs all follow similar learning trajectories in a low-dimensional InPCA space, which are distinct from those of traditional density estimation methods like histograms and Gaussian kernel density estimation (KDE). We interpret the LLaMA in-context DE process as a KDE with an adaptive kernel width and shape. This custom kernel model captures a significant portion of LLaMA's

behavior despite having only two parameters. We further speculate on why LLaMA's kernel width and shape differs from classical algorithms, providing insights into the mechanism of in-context probabilistic reasoning in LLMs.Our codebase, along with a 3D visualization of an LLM's in-context learning trajectory, is publicly available at https://github.com/AntonioLiu97/LLMICL_inPCA.

## 2444. NEAR: A Training-Free Pre-Estimator of Machine Learning Model Performance

链接：https://iclr.cc/virtual/2025/poster/29235 abstract： Artificial neural networks have been shown to be state-of-the-art machine learning models in a wide variety of applications, including natural language processing and image recognition. However, building a performant neural network is a laborious task and requires substantial computing power. Neural Architecture Search (NAS) addresses this issue by an automatic selection of the optimal network from a set of potential candidates. While many NAS methods still require training of (some) neural networks, zero-cost proxies promise to identify the optimal network without training. In this work, we propose the zero-cost proxy Network Expressivity by Activation Rank (NEAR). It is based on the effective rank of the pre- and post-activation matrix, i.e., the values of a neural network layer before and after applying its activation function. We demonstrate the cutting-edge correlation between this network score and the model accuracy on NAS-Bench-101 and NATS-Bench-SSS/TSS. In addition, we present a simple approach to estimate the optimal layer sizes in multi-layer perceptrons. Furthermore, we show that this score can be utilized to select hyperparameters such as the activation function and the neural network weight initialization scheme.

## 2445. Training-Free Diffusion Model Alignment with Sampling Demons

链接：https://iclr.cc/virtual/2025/poster/28034 abstract： Aligning diffusion models with user preferences has been a key challenge.Existing methods for aligning diffusion models either require retraining or are limited to differentiable reward functions.To address these limitations, we propose a stochastic optimization approach, dubbed Demon, to guide the denoising process at inference time without backpropagation through reward functions or model retraining.Our approach works by controlling noise distribution in denoising steps to concentrate density on regions corresponding to high rewards through stochastic optimization.We provide comprehensive theoretical and empirical evidence to support and validate our approach, including experiments that use non-differentiable sources of rewards such as Visual-Language Model (VLM) APIs and human judgements.To the best of our knowledge, the proposed approach is the first inference-time, backpropagation-free preference alignment method for diffusion models.Our method can be easily integrated with existing diffusion models without further training.Our experiments show that the proposed approach significantly improves the average aesthetics scores for text-to-image generation. Implementation is available at this URL.

## 2446. The Semantic Hub Hypothesis: Language Models Share Semantic Representations Across Languages and Modalities

链接：https://iclr.cc/virtual/2025/poster/30324 abstract： Modern language models can process inputs across diverse languages and modalities. We hypothesize that models acquire this capability through learning a shared representation space across heterogeneous data types (e.g., different languages and modalities), which places semantically similar inputs near one another, even if they are from different modalities/languages. We term this the semantic hub hypothesis, following the hub-and-spoke model from neuroscience (Patterson et al., 2007) which posits that semantic knowledge in the human brain is organized through a transmodal semantic "hub" which integrates information from various modality-specific ``spokes" regions. We first show that model representations for semantically equivalent inputs in different languages are similar in the intermediate layers, and that this space can be interpreted using the model's dominant pretraining language via the logit lens. This tendency extends to other data types, including arithmetic expressions, code, and visual/audio inputs. Interventions in the shared representation space in one data type also predictably affect model outputs in other data types, suggesting that this shared representations space is not simply a vestigial byproduct of large-scale training on broad data, but something that is actively utilized by the model during input processing.

## 2447. Towards Certification of Uncertainty Calibration under Adversarial Attacks

链接：https://iclr.cc/virtual/2025/poster/27949 abstract： Since neural classifiers are known to be sensitive to adversarial perturbations that alter their accuracy, certification methods have been developed to provide provable guarantees on the insensitivity of their predictions to such perturbations. On the other hand, in safety-critical applications, the frequentist interpretation of the confidence of a classifier (also known as model calibration) can be of utmost importance. This property can be measured via the Brier score or the expected calibration error. We show that attacks can significantly harm calibration, and thus propose certified calibration providing worst-case bounds on calibration under adversarial perturbations. Specifically, we produce analytic bounds for the Brier score and approximate bounds via the solution of a mixed-integer program on the expected calibration error. Finally, we propose novel calibration attacks and demonstrate how they can improve model calibration through adversarial calibration training. The code will be publicly released upon acceptance.

## 2448. Shh, don't say that! Domain Certification in LLMs

链接：https://iclr.cc/virtual/2025/poster/30364 abstract： Large language models (LLMs) are often deployed to do constrained tasks, with narrow domains. For example, customer support bots can be built on top of LLMs, relying on their broad language understanding and capabilities to enhance performance. However, these LLMs are adversarially susceptible, potentially generating outputs outside the intended domain. To formalize, assess and mitigate this risk, we introduce domain certification; a guarantee that accurately characterizes the out-of-domain behavior of language models. We then propose a simple yet effective approach dubbed VALID that provides adversarial bounds as a certificate. Finally, we evaluate our method across a diverse set of datasets, demonstrating that it yields meaningful certificates.

## 2449. Revisiting Feature Prediction for Learning Visual Representations from Video

链接：https://iclr.cc/virtual/2025/poster/31477 abstract： This paper explores feature prediction as a stand-alone objective for unsupervised learning from video and introduces V-JEPA, a collection of vision models trained solely using a feature prediction objective, without the use of pretrained image encoders, text, negative examples, reconstruction, or other sources of supervision. The models are trained on 2 million videos collected from public datasets and are evaluated on downstream image and video tasks. Our results show that learning by predicting video features leads to versatile visual representations that perform well on both motion and appearance-based tasks, without adaption of the model’s parameters; e.g., using a frozen backbone. Our largest model, a ViT-H/16 trained only on videos, obtains 81.9% on Kinetics-400, 72.2% on Something-Something-v2, and 77.9% on ImageNet1K.

## 2450. Simulating Human-like Daily Activities with Desire-driven Autonomy

链接：https://iclr.cc/virtual/2025/poster/31056 abstract： Desires motivate humans to interact autonomously with the complex world. In contrast, current AI agents require explicit task specifications, such as instructions or reward functions, which constrain their autonomy and behavioral diversity. In this paper, we introduce a Desire-driven Autonomous Agent (D2A) that can enable a large language model (LLM) to autonomously propose and select tasks, motivated by satisfying its multi-dimensional desires. Specifically, the motivational framework of D2A is mainly constructed by a dynamic $Value\ System$, inspired by the Theory of Needs. It incorporates an understanding of human-like desires, such as the need for social interaction, personal fulfillment, and self-care. At each step, the agent evaluates the value of its current state, proposes a set of candidate activities, and selects the one that best aligns with its intrinsic motivations. We conduct experiments on Concordia, a text-based simulator, to demonstrate that our agent generates coherent, contextually relevant daily activities while exhibiting variability and adaptability similar to human behavior. A comparative analysis with other LLM-based agents demonstrates that our approach significantly enhances the rationality of the simulated activities.

## 2451. Newton Meets Marchenko-Pastur: Massively Parallel Second-Order Optimization with Hessian Sketching and Debiasing

链接：https://iclr.cc/virtual/2025/poster/29499 abstract： Motivated by recent advances in serverless cloud computing, in particular the ``function as a service'' (FaaS) model, we consider the problem of minimizing a convex function in a massively parallel fashion, where communication between workers is limited.Focusing on the case of a twice-differentiable objective subject to an L2 penalty, we propose a scheme where the central node (server) effectively runs a Newton method, offloading its high per-iteration cost---stemming from the need to invert the Hessian---to the workers. In our solution, workers produce independently coarse but low-bias estimates of the inverse Hessian, using an adaptive sketching scheme. The server then averages the descent directions produced by the workers, yielding a good approximation for the exact Newton step. The main component of our adaptive sketching scheme is a low-complexity procedure for selecting the sketching dimension, an issue that was left largely unaddressed in the existing literature on Hessian sketching for distributed optimization. Our solution is based on ideas from asymptotic random matrix theory, specifically the Marchenko-Pastur law. For Gaussian sketching matrices, we derive non asymptotic guarantees for our algorithm which do not depend on the condition number of the Hessian nor a priori require the sketching dimension to be proportional to the dimension, as is often the case in asymptotic random matrix theory. Lastly, when the objective is self-concordant, we provide convergence guarantees for the approximate Newton's method with noisy Hessians, which may be of independent interest beyond the setting considered in this paper.

## 2452. Theory, Analysis, and Best Practices for Sigmoid Self-Attention

链接：https://iclr.cc/virtual/2025/poster/29205 abstract： Attention is a key part of the transformer architecture. It is a sequence-to-sequence mapping that transforms each sequence element into a weighted sum of values. The weights are typically obtained as the softmax of dot products between keys and queries. Recent work has explored alternatives to softmax attention in transformers, such as ReLU and sigmoid activations. In this work, we revisit sigmoid attention and conduct an in-depth theoretical and empirical analysis. Theoretically, we prove that transformers with sigmoid attention are universal function approximators and benefit from improved regularity compared to softmax attention. Through detailed empirical analysis, we identify stabilization of large initial attention norms during the early stages of training as a crucial factor for the successful training of models with sigmoid attention, outperforming prior attempts. We also introduce FLASHSIGMOID, a hardware-aware and memory-efficient implementation of sigmoid attention yielding a 17% inference kernel speed-up over FLASHATTENTION2 on H100 GPUs. Experiments across language, vision, and speech show that properly normalized sigmoid attention matches the strong performance of softmax attention on a wide range of domains and scales, which previous attempts at sigmoid attention

were unable to fully achieve. Our work unifies prior art and establishes best practices for sigmoid attention as a drop-in softmax replacement in transformers.

## 2453. Uncertainty-Aware Decoding with Minimum Bayes Risk

链接：https://iclr.cc/virtual/2025/poster/28763 abstract： Despite their outstanding performance in the majority of scenarios, contemporary language models still occasionally generate undesirable outputs, for example, hallucinated text. While such behaviors have previously been linked to uncertainty, there is a notable lack of methods that actively consider uncertainty during text generation. In this work, we show how Minimum Bayes Risk (MBR) decoding, which selects model generations according to an expected risk, can be generalized into a principled uncertainty-aware decoding method. In short, we account for model uncertainty during decoding by incorporating a posterior over model parameters into MBR's computation of expected risk. We show that this modified expected risk is useful for both choosing outputs and deciding when to abstain from generation and can provide improvements without incurring overhead. We benchmark different methods for learning posteriors and show that performance improves with prediction diversity. We release our code publicly.

## 2454. TopoNets: High performing vision and language models with brain-like topography

链接：https://iclr.cc/virtual/2025/poster/29552 abstract： Neurons in the brain are organized such that nearby cells tend to share similar functions. AI models lack this organization, and past efforts to introduce topography have often led to trade-offs between topography and task performance. In this work, we present TopoLoss, a new loss function that promotes spatially organized topographic representations in AI models without significantly sacrificing task performance. TopoLoss is highly adaptable and can be seamlessly integrated into the training of leading model architectures. We validate our method on both vision (ResNet-18, ResNet-50, ViT) and language models (GPT-Neo-125M, NanoGPT), collectively TopoNets. TopoNets are the highest performing supervised topographic models to date, exhibiting brain-like properties such as localized feature processing, lower dimensionality, and increased efficiency. TopoNets also predict responses in the brain and replicate the key topographic signatures observed in the brain's visual and language cortices, further bridging the gap between biological and artificial systems. This work establishes a robust and generalizable framework for integrating topography into AI, advancing the development of high performing models that more closely emulate the computational strategies of the human brain. Our project page: https://toponets.github.io

## 2455. OptiBench Meets ReSocratic: Measure and Improve LLMs for Optimization Modeling

链接：https://iclr.cc/virtual/2025/poster/28850 abstract： Large language models (LLMs) have exhibited their problem-solving abilities in mathematical reasoning. Solving realistic optimization (OPT) problems in application scenarios requires advanced and applied mathematics ability. However, current OPT benchmarks that merely solve linear programming are far from complex realistic situations. In this work, we propose OptiBench, a benchmark for End-to-end optimization problem-solving with human-readable inputs and outputs. OptiBench contains rich optimization problems, including linear and nonlinear programming with or without tabular data, which can comprehensively evaluate LLMs' solving ability. In our benchmark, LLMs are required to call a code solver to provide precise numerical answers.Furthermore, to alleviate the data scarcity for optimization problems, and to bridge the gap between open-source LLMs on a small scale (e.g., Llama-3-8b) and closed-source LLMs (e.g., GPT-4), we further propose a data synthesis method namely ReSocratic. Unlike general data synthesis methods that proceed from questions to answers, \ReSocratic first incrementally synthesizes formatted optimization demonstration with mathematical formulations step by step and then back-translates the generated demonstrations into questions. Based on this, we synthesize the ReSocratic-29k dataset. We further conduct supervised fine-tuning with ReSocratic-29k on multiple open-source models. Experimental results show that ReSocratic-29k significantly improves the performance of open-source models.

## 2456. Hierarchical World Models as Visual Whole-Body Humanoid Controllers

链接：https://iclr.cc/virtual/2025/poster/30793 abstract： Whole-body control for humanoids is challenging due to the high-dimensional nature of the problem, coupled with the inherent instability of a bipedal morphology. Learning from visual observations further exacerbates this difficulty. In this work, we explore highly data-driven approaches to visual whole-body humanoid control based on reinforcement learning, without any simplifying assumptions, reward design, or skill primitives. Specifically, we propose a hierarchical world model in which a high-level agent generates commands based on visual observations for a low-level agent to execute, both of which are trained with rewards. Our approach produces highly performant control policies in 8 tasks with a simulated 56-DoF humanoid, while synthesizing motions that are broadly preferred by humans. Code and videos: https://www.nicklashansen.com/rlpuppeteer

## 2457. Seq-VCR: Preventing Collapse in Intermediate Transformer Representations for Enhanced Reasoning

链接：https://iclr.cc/virtual/2025/poster/31097 abstract： Decoder-only Transformers often struggle with complex reasoning tasks, particularly arithmetic reasoning requiring multiple sequential operations. In this work, we identify representation collapse in the model's intermediate layers as a key factor limiting their reasoning capabilities. To address this, we propose Sequential Variance-Covariance Regularization (Seq-VCR), which enhances the entropy of intermediate representations and prevents collapse. Combined with dummy pause tokens as substitutes for chain-of-thought (CoT) tokens, our method significantly improves performance in arithmetic reasoning problems. In the challenging 5 × 5 integer multiplication task, our approach achieves 99.5% exact match accuracy, outperforming models of the same size (which yield 0% accuracy) and GPT-4 with five-shot CoT prompting (44%). We also demonstrate superior results on arithmetic expression and longest increasing subsequence (LIS) datasets. Our findings highlight the importance of preventing intermediate layer representation collapse to enhance the reasoning capabilities of Transformers and show that Seq-VCR offers an effective solution without requiring explicit CoT supervision.

## 2458. FormalAlign: Automated Alignment Evaluation for Autoformalization

链接：https://iclr.cc/virtual/2025/poster/30594 abstract： Autoformalization aims to convert informal mathematical proofs into machine-verifiable formats, bridging the gap between natural and formal languages. However, ensuring semantic alignment between the informal and formalized statements remains challenging. Existing approaches heavily rely on manual verification, hindering scalability. To address this, we introduce FormalAlign, a framework for automatically evaluating the alignment between natural and formal languages in autoformalization. FormalAlign trains on both the autoformalization sequence generation task and the representational alignment between input and output, employing a dual loss that combines a pair of mutually enhancing autoformalization and alignment tasks. Evaluated across four benchmarks augmented by our proposed misalignment strategies, FormalAlign demonstrates superior performance. In our experiments, FormalAlign outperforms GPT-4, achieving an Alignment-Selection Score 11.58\% higher on \forml-Basic (99.21\% vs. 88.91\%) and 3.19\% higher on MiniF2F-Valid (66.39\% vs. 64.34\%). This effective alignment evaluation significantly reduces the need for manual verification.

## 2459. LDAdam: Adaptive Optimization from Low-Dimensional Gradient Statistics

链接：https://iclr.cc/virtual/2025/poster/29199 abstract： We introduce LDAdam, a memory-efficient optimizer for training large models, that performs adaptive optimization steps within lower dimensional subspaces, while consistently exploring the full parameter space during training. This strategy keeps the optimizer's memory footprint to a fraction of the model size. LDAdam relies on a new projection-aware update rule for the optimizer states that allows for transitioning between subspaces, i.e., estimation of the statistics of the projected gradients. To mitigate the errors due to low-rank projection, LDAdam integrates a new generalized error feedback mechanism, which explicitly accounts for both gradient and optimizer state compression. We prove the convergence of LDAdam under standard assumptions, and provide empirical evidence that LDAdam allows for efficient fine-tuning and pre-training of language models.

## 2460. Regressing the Relative Future: Efficient Policy Optimization for Multi-turn RLHF

链接：https://iclr.cc/virtual/2025/poster/29044 abstract： Large Language Models (LLMs) have achieved remarkable success at tasks like summarization that involve a single turn of interaction. However, they can still struggle with multi-turn tasks like dialogue that require long-term planning. Previous works on multi-turn dialogue extend single-turn reinforcement learning from human feedback (RLHF) methods to the multi-turn setting by treating all prior dialogue turns as a long context. Such approaches suffer from covariate shift: the conversations in the training set have previous turns generated by some reference policy, which means that low training error may not necessarily correspond to good performance when the learner is actually in the conversation loop. In response, we introduce REgressing the RELative FUture (REFUEL), an efficient policy optimization approach designed to address multi-turn RLHF in LLMs. REFUEL employs a single model to estimate $Q$-values and trains on self-generated data, addressing the covariate shift issue. REFUEL frames the multi-turn RLHF problem as a sequence of regression tasks on iteratively collected datasets, enabling ease of implementation. Theoretically, we prove that REFUEL can match the performance of any policy covered by the training set. Empirically, we evaluate our algorithm by using Llama-3.1-70B-it to simulate a user in conversation with our model. REFUEL consistently outperforms state-of-the-art methods such as DPO and REBEL across various settings. Furthermore, despite having only 8 billion parameters, Llama-3-8B-it fine-tuned with REFUEL outperforms Llama-3.1-70B-it on long multi-turn dialogues. Implementation of REFUEL can be found at https://github.com/ZhaolinGao/REFUEL/, and models trained by REFUEL can be found at https://huggingface.co/Cornell-AGI.

## 2461. PharmacoMatch: Efficient 3D Pharmacophore Screening via Neural Subgraph Matching

链接：https://iclr.cc/virtual/2025/poster/31161 abstract： The increasing size of screening libraries poses a significant challenge for the development of virtual screening methods for drug discovery, necessitating a re-evaluation of traditional approaches in the era of big data. Although 3D pharmacophore screening remains a prevalent technique, its application to very large datasets is limited by the computational cost associated with matching query pharmacophores to database molecules. In this study, we introduce PharmacoMatch, a novel contrastive learning approach based on neural subgraph matching. Our

method reinterprets pharmacophore screening as an approximate subgraph matching problem and enables efficient querying of conformational databases by encoding query-target relationships in the embedding space. We conduct comprehensive investigations of the learned representations and evaluate PharmacoMatch as pre-screening tool in a zero-shot setting. We demonstrate significantly shorter runtimes and comparable performance metrics to existing solutions, providing a promising speed-up for screening very large datasets.

## 2462. Tracking objects that change in appearance with phase synchrony

链接：https://iclr.cc/virtual/2025/poster/28488 abstract： Objects we encounter often change appearance as we interact with them. Changes in illumination (shadows), object pose, or the movement of non-rigid objects can drastically alter available image features. How do biological visual systems track objects as they change? One plausible mechanism involves attentional mechanisms for reasoning about the locations of objects independently of their appearances --- a capability that prominent neuroscience theories have associated with computing through neural synchrony. Here, we describe a novel deep learning circuit that can learn to precisely control attention to features separately from their location in the world through neural synchrony: the complex-valued recurrent neural network (CV-RNN). Next, we compare object tracking in humans, the CV-RNN, and other deep neural networks (DNNs), using FeatureTracker: a large-scale challenge that asks observers to track objects as their locations and appearances change in precisely controlled ways. While humans effortlessly solved FeatureTracker, state-of-the-art DNNs did not. In contrast, our CV-RNN behaved similarly to humans on the challenge, providing a computational proof-of-concept for the role of phase synchronization as a neural substrate for tracking appearance-morphing objects as they move about.

## 2463. InversionGNN: A Dual Path Network for Multi-Property Molecular Optimization

链接：https://iclr.cc/virtual/2025/poster/28405 abstract： Exploring chemical space to find novel molecules that simultaneously satisfy multiple properties is crucial in drug discovery. However, existing methods often struggle with trading off multiple properties due to the conflicting or correlated nature of chemical properties. To tackle this issue, we introduce InversionGNN framework, an effective yet sample-efficient dual-path graph neural network (GNN) for multi-objective drug discovery. In the direct prediction path of InversionGNN, we train the model for multi-property prediction to acquire knowledge of the optimal combination of functional groups.Then the learned chemical knowledge helps the inversion generation path to generate molecules with required properties. In order to decode the complex knowledge of multiple properties in the inversion path, we propose a gradient-based Pareto search method to balance conflicting properties and generate Pareto optimal molecules. Additionally, InversionGNN is able to search the full Pareto front approximately in discrete chemical space. Comprehensive experimental evaluations show that InversionGNN is both effective and sample-efficient in various discrete multi-objective settings including drug discovery.

## 2464. Turning Up the Heat: Min-p Sampling for Creative and Coherent LLM Outputs

链接：https://iclr.cc/virtual/2025/poster/30358 abstract： Large Language Models (LLMs) generate text by sampling the next token from a probability distribution over the vocabulary at each decoding step. Popular sampling methods like top-p (nucleus sampling) often struggle to balance quality and diversity, especially at higher temperatures which lead to incoherent or repetitive outputs. We propose min-p sampling, a dynamic truncation method that adjusts the sampling threshold based on the model's confidence by using the top token's probability as a scaling factor. Our experiments on benchmarks including GPQA, GSM8K, and AlpacaEval Creative Writing show that min-p sampling improves both the quality and diversity of generated text across different model families (Mistral and Llama 3) and model sizes (1B to 123B parameters), especially at higher temperatures. Human evaluations further show a clear preference for min-p sampling, in both text quality and creativity. Min-p sampling has been adopted by popular open-source LLM frameworks, including Hugging Face Transformers, VLLM, and many others, highlighting its significant impact on improving text generation quality.

## 2465. CoInD: Enabling Logical Compositions in Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29065 abstract： How can we learn generative models to sample data with arbitrary logical compositions of statistically independent attributes? The prevailing solution is to sample from distributions expressed as a composition of attributes' conditional marginal distributions under the assumption that they are statistically independent. This paper shows that standard conditional diffusion models violate this assumption, even when all attribute compositions are observed during training. And, this violation is significantly more severe when only a subset of the compositions is observed. We propose CoInD to address this problem. It explicitly enforces statistical independence between the conditional marginal distributions by minimizing Fisher’s divergence between the joint and marginal distributions. The theoretical advantages of CoInD are reflected in both qualitative and quantitative experiments, demonstrating a significantly more faithful and controlled generation of samples for arbitrary logical compositions of attributes. The benefit is more pronounced for scenarios that current solutions relying on the assumption of conditionally independent marginals struggle with, namely, logical compositions involving the NOT operation and when only a subset of compositions are observed during training.

## 2466. Descent with Misaligned Gradients and Applications to Hidden Convexity

链接：https://iclr.cc/virtual/2025/poster/31146 abstract： We consider the problem of minimizing a convex objective given access to an oracle that outputs "misaligned" stochastic gradients, where the expected value of the output is guaranteed to be correlated with, but not necessarily equal to the true gradient of the objective. In the case where the misalignment (or bias) of the oracle changes slowly, we obtain an optimization algorithm that achieves the optimum iteration complexity of $\tilde O(\epsilon^{-2})$; for the more general case where the changes need not be slow, we obtain an algorithm with $\tilde O(\epsilon^{-3})$ iteration complexity. As an application of our framework, we consider optimization problems with a "hidden convexity" property, and obtain an algorithm with $O(\epsilon^{-3})$ iteration complexity.

## 2467. OASIS Uncovers: High-Quality T2I Models, Same Old Stereotypes

链接：https://iclr.cc/virtual/2025/poster/32095 abstract： Images generated by text-to-image (T2I) models often exhibit visual biases and stereotypes of concepts such as culture and profession. Existing quantitative measures of stereotypes are based on statistical parity that does not align with the sociological definition of stereotypes and, therefore, incorrectly categorizes biases as stereotypes. Instead of oversimplifying stereotypes as biases, we propose a quantitative measure of stereotypes that aligns with its sociological definition. We then propose OASIS to measure the stereotypes in a generated dataset and understand their origins within the T2I model. OASIS includes two scores to measure stereotypes from a generated image dataset: (M1) Stereotype Score to measure the distributional violation of stereotypical attributes, and (M2) WALS to measure spectral variance in the images along a stereotypical attribute. OASISalso includes two methods to understand the origins of stereotypes in T2I models: (U1) StOP to discover attributes that the T2I model internally associates with a given concept, and (U2) SPI to quantify the emergence of stereotypical attributes in the latent space of the T2I model during image generation. Despite the considerable progress in image fidelity, using OASIS, we conclude that newer T2I models such as FLUX.1 and SDv3 contain strong stereotypical predispositions about concepts and still generate images with widespread stereotypical attributes. Additionally, the quantity of stereotypes worsens for nationalities with lower Internet footprints.

## 2468. Unlearn and Burn: Adversarial Machine Unlearning Requests Destroy Model Accuracy

链接：https://iclr.cc/virtual/2025/poster/30919 abstract： Machine unlearning algorithms, designed for selective removal of training data from models, have emerged as a promising approach to growing privacy concerns. In this work, we expose a critical yet underexplored vulnerability in the deployment of unlearning systems: the assumption that the data requested for removal is always part of the original training set. We present a threat model where an attacker can degrade model accuracy by submitting adversarial unlearning requests for data \textit{not} present in the training set. We propose white-box and black-box attack algorithms and evaluate them through a case study on image classification tasks using the CIFAR-10 and ImageNet datasets, targeting a family of widely used unlearning methods. Our results show extremely poor test accuracy following the attack—3.6% on CIFAR-10 and 0.4% on ImageNet for white-box attacks, and 8.5% on CIFAR-10 and 1.3% on ImageNet for black-box attacks. Additionally, we evaluate various verification mechanisms to detect the legitimacy of unlearning requests and reveal the challenges in verification, as most of the mechanisms fail to detect stealthy attacks without severely impairing their ability to process valid requests. These findings underscore the urgent need for research on more robust request verification methods and unlearning protocols, should the deployment of machine unlearning systems become more prevalent in the future.

## 2469. $\mathbb{X}$-Sample Contrastive Loss: Improving Contrastive Learning with Sample Similarity Graphs

链接：https://iclr.cc/virtual/2025/poster/29076 abstract： Learning good representations involves capturing the diverse ways in which data samples relate. Contrastive loss—an objective matching related samples—underlies methods from self-supervised to multimodal learning. Contrastive losses, however, can be viewed more broadly as modifying a similarity graph to indicate how samples should relate in the embedding space. This view reveals a shortcoming in contrastive learning: the similarity graph is binary, as only one sample is the related positive sample. Crucially, similarities \textit{across} samples are ignored. Based on this observation, we revise the standard contrastive loss to explicitly encode how a sample relates to others. We experiment with this new objective, called $\mathbb{X}$-Sample Contrastive, to train vision models based on similarities in class or text caption descriptions.Our study spans three scales: ImageNet-1k with 1 million, CC3M with 3 million, and CC12M with 12 million samples. The representations learned via our objective outperform both contrastive self-supervised and vision-language models trained on the same data across a range of tasks. When training on CC12M, we outperform CLIP by $0.6\%$ on both ImageNet and ImageNet Real. Our objective appears to work particularly well in lower-data regimes, with gains over CLIP of $17.2\%$ on ImageNet and $18.0\%$ on ImageNet Real when training with CC3M. Finally, our objective encourages the model to learn representations that separate objects from their attributes and backgrounds, with gains of $3.3$-$5.6$\% over CLIP on ImageNet9. The proposed method takes a step towards developing richer learning objectives for understanding sample relations in foundation models.

## 2470. Diffusion State-Guided Projected Gradient for Inverse Problems

链接：https://iclr.cc/virtual/2025/poster/28584 abstract： Recent advancements in diffusion models have been effective in learning data priors for solving inverse problems. They leverage diffusion sampling steps for inducing a data prior while using a measurement guidance gradient at each step to impose data consistency. For general inverse problems, approximations are needed when an unconditionally trained diffusion model is used since the measurement likelihood is intractable, leading to inaccurate posterior sampling. In other words, due to their approximations, these methods fail to preserve the generation process on the data manifold defined by the diffusion prior, leading to artifacts in applications such as image restoration. To enhance the performance and robustness of diffusion models in solving inverse problems, we propose Diffusion State-Guided Projected Gradient (DiffStateGrad), which projects the measurement gradient onto a subspace that is a low-rank approximation of an intermediate state of the diffusion process. DiffStateGrad, as a module, can be added to a wide range of diffusion-based inverse solvers to improve the preservation of the diffusion process on the prior manifold and filter out artifact-inducing components. We highlight that DiffStateGrad improves the robustness of diffusion models in terms of the choice of measurement guidance step size and noise while improving the worst-case performance. Finally, we demonstrate that DiffStateGrad improves upon the state-of-the-art on linear and nonlinear image restoration inverse problems. Our code is available at https://github.com/Anima-Lab/DiffStateGrad.

## 2471. LiveBench: A Challenging, Contamination-Limited LLM Benchmark

链接：https://iclr.cc/virtual/2025/poster/28134 abstract： Test set contamination, wherein test data from a benchmark ends up in a newer model's training set, is a well-documented obstacle for fair LLM evaluation and can quickly render benchmarks obsolete. To mitigate this, many recent benchmarks crowdsource new prompts and evaluations from human or LLM judges; however, these can introduce significant biases, and break down when scoring hard questions. In this work, we introduce a new benchmark for LLMs designed to be resistant to both test set contamination and the pitfalls of LLM judging and human crowdsourcing. We release LiveBench, the first benchmark that (1) contains frequently-updated questions from recent information sources, (2) scores answers automatically according to objective ground-truth values, and (3) contains a wide variety of challenging tasks, spanning math, coding, reasoning, language, instruction following, and data analysis. To achieve this, LiveBench contains questions that are based on recently-released math competitions, arXiv papers, news articles, and datasets, and it contains harder, contamination-limited versions of tasks from previous benchmarks such as Big-Bench Hard, AMPS, and IFEval. We evaluate many prominent closed-source models, as well as dozens of open-source models ranging from 0.5B to 405B in size. LiveBench is difficult, with top models achieving below 70% accuracy. We release all questions, code, and model answers. Questions are added and updated on a monthly basis, and we release new tasks and harder versions of tasks over time so that LiveBench can distinguish between the capabilities of LLMs as they improve in the future. We welcome community engagement and collaboration for expanding the benchmark tasks and models.

## 2472. Filtered not Mixed: Filtering-Based Online Gating for Mixture of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28916 abstract： We propose MoE-F — a formalized mechanism for combining N pre-trained expert Large Language Models (LLMs) in online time-series prediction tasks by adaptively forecasting the best weighting of LLM predictions at every time step. Our mechanism leverages the conditional information in each expert's running performance to forecast the best combination of LLMs for predicting the time series in its next step. Diverging from static (learned) Mixture of Experts (MoE) methods, our approach employs time-adaptive stochastic filtering techniques to combine experts. By framing the expert selection problem as a finite state-space, continuous-time Hidden Markov model (HMM), we can leverage the Wohman-Shiryaev filter. Our approach first constructs N parallel filters corresponding to each of the N individual LLMs. Each filter proposes its best combination of LLMs, given the information that they have access to. Subsequently, the N filter outputs are optimally aggregated to maximize their robust predictive power, and this update is computed efficiently via a closed-form expression, thus generating our ensemble predictor.Our contributions are:- (I) the MoE-F algorithm — deployable as a plug-and-play filtering harness,- (II) theoretical optimality guarantees of the proposed filtering-based gating algorithm (via optimality guarantees for its parallel Bayesian filtering and its robust aggregation steps), and- (III) empirical evaluation and ablative results using state-of-the-art foundational and MoE LLMs on a real-world Financial Market Movement task where MoE-F attains a remarkable 17% absolute and 48.5% relative F1 measure improvement over the next best performing individual LLM expert predicting short-horizon market movement based on streaming news. Further, we provide empirical evidence of substantial performance gains in applying MoE-F over specialized models in the long-horizon time-series forecasting domain. Code available on github: https://github.com/raeidsaqur/moe-f

## 2473. API Pack: A Massive Multi-Programming Language Dataset for API Call Generation

链接：https://iclr.cc/virtual/2025/poster/28893 abstract： We introduce API Pack, a massive multi-programming language dataset containing over one million instruction-API calls for improving the API call generation capabilities of large language models. Our evaluation highlights three key findings: First, fine-tuning on API Pack enables open-source models to outperform GPT-3.5 and GPT-4 in generating code for entirely new API calls. We show this by fine-tuning CodeLlama-13B on 20,000 Python instances from API Pack. Second, fine-tuning on a large dataset in one language, combined with smaller datasets from others, improves API generation accuracy across multiple languages. Third, we confirm the benefits of larger datasets for API generalization, as increasing fine-tuning data to one million instances enhances generalization to new APIs. To support further research, we open-source the API Pack dataset, trained model, and code at https://github.com/zguo0525/API-Pack.

## 2474. Learning from weak labelers as constraints

链接：https://iclr.cc/virtual/2025/poster/31155 abstract： We study programmatic weak supervision, where in contrast to labeled data, we have access to \emph{weak labelers}, each of which either abstains or provides noisy labels corresponding to any input. Most previous approaches typically employ latent generative models that model the joint distribution of the weak labels and the latent ``true'' label. The caveats are that this relies on assumptions that may not always hold in practice such as conditional independence assumptions over the joint distribution of the weak labelers and the latent true label, and more general implicit inductive biases in the latent generative models. In this work, we consider a more explicit form of side-information that can be leveraged to denoise the weak labeler, namely the bounds on the average error of the weak labelers. We then propose a novel but natural weak supervision objective that minimizes a regularization functional subject to satisfying these bounds. This turns out to be a difficult constrained optimization problem due to discontinuous accuracy bound constraints. We provide a continuous optimization formulation for this objective through an alternating minimization algorithm that iteratively computes soft pseudo labels on the unlabeled data satisfying the constraints while being close to the model, and then updates the model on these labels until all the constraints are satisfied. We follow this with a theoretical analysis of this approach and provide insights into its denoising effects in training discriminative models given multiple weak labelers. Finally, we demonstrate the superior performance and robustness of our method on a popular weak supervision benchmark.

## 2475. Hummingbird: High Fidelity Image Generation via Multimodal Context Alignment

链接：https://iclr.cc/virtual/2025/poster/30868 abstract： While diffusion models are powerful in generating high-quality, diverse synthetic data for object-centric tasks, existing methods struggle with scene-aware tasks such as Visual Question Answering (VQA) and Human-Object Interaction (HOI) Reasoning, where it is critical to preserve scene attributes in generated images consistent with a multimodal context, i.e. a reference image with accompanying text guidance query. To address this, we introduce Hummingbird, the first diffusion-based image generator which, given a multimodal context, generates highly diverse images w.r.t. the reference image while ensuring high fidelity by accurately preserving scene attributes, such as object interactions and spatial relationships from the text guidance. Hummingbird employs a novel Multimodal Context Evaluator that simultaneously optimizes our formulated Global Semantic and Fine-grained Consistency Rewards to ensure generated images preserve the scene attributes of reference images in relation to the text guidance while maintaining diversity. As the first model to address the task of maintaining both diversity and fidelity given a multimodal context, we introduce a new benchmark formulation incorporating MME Perception and Bongard HOI datasets. Benchmark experiments show Hummingbird outperforms all existing methods by achieving superior fidelity while maintaining diversity, validating Hummingbird's potential as a robust multimodal context-aligned image generator in complex visual tasks. Project page: https://roar-ai.github.io/hummingbird

## 2476. Neuron Platonic Intrinsic Representation From Dynamics Using Contrastive Learning

链接：https://iclr.cc/virtual/2025/poster/27925 abstract： The Platonic Representation Hypothesis posits that behind different modalities of data (what we sense or detect), there exists a universal, modality-independent representation of reality. Inspired by this, we treat each neuron as a system, where we can detect the neuron’s multi-segment activity data under different peripheral conditions. We believe that, similar to the Platonic idea, there exists a time-invariant representation behind the different segments of the same neuron, which reflects the intrinsic properties of the neuron’s system. Intrinsic properties include the molecular profiles, brain regions and morphological structure, etc. The optimization objective for obtaining the intrinsic representation of neurons should satisfy two criteria: (I) segments from the same neuron should have a higher similarity than segments from different neurons; (II) the representations should generalize well to out-of-domain data. To achieve this, we employ contrastive learning, treating different segments from the same neuron as positive pairs and segments from different neurons as negative pairs. During the implementation, we chose the VICReg, which uses only positive pairs for optimization but indirectly separates dissimilar samples via regularization terms. To validate the efficacy of our method, we first applied it to simulated neuron population dynamics data generated using the Izhikevich model. We successfully confirmed that our approach captures the type of each neuron as defined by preset hyperparameters. We then applied our method to two real-world neuron dynamics datasets, including spatial transcriptomics-derived neuron type annotations and the brain regions where each neuron is located. The learned representations from our model not only predict neuron type and location but also show robustness when tested on out-of-domain data (unseen animals). This demonstrates the potential of our approach in advancing the understanding of neuronal systems and offers valuable insights for future neuroscience research.

## 2477. Instant Policy: In-Context Imitation Learning via Graph Diffusion

链接：https://iclr.cc/virtual/2025/poster/28628 abstract： Following the impressive capabilities of in-context learning with large transformers, In-Context Imitation Learning (ICIL) is a promising opportunity for robotics. We introduce Instant Policy, which learns new tasks instantly from just one or two demonstrations, achieving ICIL through two key components. First, we introduce inductive biases through a graph representation and model ICIL as a graph generation problem using a learned diffusion process, enabling structured reasoning over demonstrations, observations, and actions. Second, we show that such a model can be trained using pseudo-demonstrations – arbitrary trajectories generated in simulation – as a virtually infinite pool of training data. Our experiments, in both simulation and reality, show that Instant Policy enables rapid learning of various everyday robot tasks. We also show how it can serve as a foundation for cross-embodiment and zero-shot transfer to language-defined

tasks.

## 2478. Group Downsampling with Equivariant Anti-aliasing

链接：https://iclr.cc/virtual/2025/poster/28130 abstract： Downsampling layers are crucial building blocks in CNN architectures, which help to increase the receptive field for learning high-level features and reduce the amount of memory/computation in the model. In this work, we study the generalization of the uniform downsampling layer for group equivariant architectures, e.g., $G$-CNNs. That is, we aim to downsample signals (feature maps) on general finite groups *with* anti-aliasing. This involves the following: **(a)** Given a finite group and a downsampling rate, we present an algorithm to form a suitable choice of subgroup. **(b)** Given a group and a subgroup, we study the notion of bandlimited-ness and propose how to perform anti-aliasing. Notably, our method generalizes the notion of downsampling based on classical sampling theory. When the signal is on a cyclic group, i.e., periodic, our method recovers the standard downsampling of an ideal low-pass filter followed by a subsampling operation. Finally, we conducted experiments on image classification tasks demonstrating that the proposed downsampling operation improves accuracy, better preserves equivariance, and reduces model size when incorporated into $G$-equivariant networks

## 2479. LoRA Learns Less and Forgets Less

链接：https://iclr.cc/virtual/2025/poster/31465 abstract： Low-Rank Adaptation (LoRA) is a widely-used parameter-efficient finetuning method for large language models. LoRA saves memory by training only low rank perturbations to selected weight matrices. In this work, we compare the performance of LoRA and full finetuning on two target domains, programming and mathematics. We consider both the instruction finetuning ($\approx$100K prompt-response pairs) and continued pretraining ($\approx$20B unstructured tokens) data regimes. Our results show that, in the standard low-rank settings, LoRA substantially underperforms full finetuning. Nevertheless, LoRA better maintains the base model's performance on tasks outside the target domain. We show that LoRA mitigates forgetting more than common regularization techniques such as weight decay and dropout; it also helps maintain more diverse generations. Finally, we show that full finetuning learns perturbations with a rank that is 10-100$\times$ greater than typical LoRA configurations, possibly explaining some of the reported gaps. We conclude by proposing best practices for finetuning with LoRA.

## 2480. A Distributional Approach to Uncertainty-Aware Preference Alignment Using Offline Demonstrations

链接：https://iclr.cc/virtual/2025/poster/29655 abstract： Designing reward functions in Reinforcement Learning (RL) often demands significant task-specific expertise. Offline Preference-based Reinforcement Learning (PbRL) provides an effective alternative to address the complexity of reward design by learning policies from offline datasets that contain human preferences between trajectory pairs. Existing offline PbRL studies typically model a reward function by maximizing its likelihood of generating the observed human preferences. However, due to the varying number of samples within the limited dataset, less frequently compared trajectories exhibit greater uncertainty, which potentially leads to unreliable behaviors during reward and policy updates. To solve this issue, in this work, we introduce Uncertainty-Aware PbRL (UA-PbRL) to learn a distributional reward model and a risk-sensitive policy from an offline preference dataset. Our approach employs a Maximum A Posteriori (MAP) objective to update trajectory rewards and incorporates an informative prior to account for the uncertainties. Building upon this reward update, we propose a generative reward model to capture the reward distribution, utilizing the offline distributional Bellman operator and the Conditional Value-at-Risk (CVaR) metric to train a risk-sensitive policy. Experimental results demonstrate that UA-PbRL effectively identifies and avoids states with high uncertainty, facilitating risk-averse behaviors across various tasks, including robot control and language model alignment. The code is available at https://github.com/Jasonxu1225/UA-PbRL.

## 2481. Estimating the Probabilities of Rare Outputs in Language Models

链接：https://iclr.cc/virtual/2025/poster/30474 abstract： We consider the problem of low probability estimation: given a machine learning model and a formally-specified input distribution, how can we estimate the probability of a binary property of the model's output, even when that probability is too small to estimate by random sampling? This problem is motivated by the need to improve worst-case performance, which distribution shift can make much more likely. We study low probability estimation in the context of argmax sampling from small transformer language models. We compare two types of methods: importance sampling, which involves searching for inputs giving rise to the rare output, and activation extrapolation, which involves extrapolating a probability distribution fit to the model's logits. We find that importance sampling outperforms activation extrapolation, but both outperform naive sampling. Finally, we explain how minimizing the probability estimate of an undesirable behavior generalizes adversarial training, and argue that new methods for low probability estimation are needed to provide stronger guarantees about worst-case performance.

## 2482. Self-Normalized Resets for Plasticity in Continual Learning

链接：https://iclr.cc/virtual/2025/poster/30297 abstract： Plasticity Loss is an increasingly important phenomenon that refers to the empirical observation that as a neural network is continually trained on a sequence of changing tasks, its ability to adapt to a new task diminishes over time. We introduce Self-Normalized Resets (SNR), a simple adaptive algorithm that mitigates plasticity loss by resetting a neuron’s weights when evidence suggests its firing rate has effectively dropped to zero. Across a

battery of continual learning problems and network architectures, we demonstrate that SNR consistently attains superior performance compared to its competitor algorithms. We also demonstrate that SNR is robust to its sole hyperparameter, its rejection percentile threshold, while competitor algorithms show significant sensitivity. SNR's threshold-based reset mechanism is motivated by a simple hypothesis test we derive. Seen through the lens of this hypothesis test, competing reset proposals yield suboptimal error rates in correctly detecting inactive neurons, potentially explaining our experimental observations. We also conduct a theoretical investigation of the optimization landscape for the problem of learning a single ReLU. We show that even when initialized adversarially, an idealized version of SNR learns the target ReLU, while regularization based approaches can fail to learn.

## 2483. DiSK: Differentially Private Optimizer with Simplified Kalman Filter for Noise Reduction

链接：https://iclr.cc/virtual/2025/poster/29981 abstract： Differential privacy (DP) offers a robust framework for safeguarding individual data privacy. To utilize DP in training modern machine learning models, differentially private optimizers have been widely used in recent years. A popular approach to privatize an optimizer is to clip the individual gradients and add sufficiently large noise to the clipped gradient. This approach led to the development of DP optimizers that have comparable performance with their non-private counterparts in fine-tuning tasks or in tasks with a small number of training parameters. However, a significant performance drop is observed when these optimizers are applied to large-scale training. This degradation stems from the substantial noise injection required to maintain DP, which disrupts the optimizer's dynamics.This paper introduces DiSK, a novel framework designed to significantly enhance the performance of DP optimizers. DiSK employs Kalman filtering, a technique drawn from control and signal processing, to effectively denoise privatized gradients and generate progressively refined gradient estimations. To ensure practicality for large-scale training, we simplify the Kalman filtering process, minimizing its memory and computational demands.We establish theoretical privacy-utility trade-off guarantees for DiSK, and demonstrate provable improvements over standard DP optimizers like DPSGD in terms of iteration complexity upper-bound.Extensive experiments across diverse tasks, including vision tasks such as CIFAR-100 and ImageNet-1k and language fine-tuning tasks such as GLUE, E2E, and DART, validate the effectiveness of DiSK. The results showcase its ability to significantly improve the performance of DP optimizers, surpassing state-of-the-art results under the same privacy constraints on several benchmarks.

## 2484. Addax: Utilizing Zeroth-Order Gradients to Improve Memory Efficiency and Performance of SGD for Fine-Tuning Language Models

链接：https://iclr.cc/virtual/2025/poster/29681 abstract： Fine-tuning language models (LMs) with the standard Adam optimizer often demands excessive memory, limiting accessibility. The ``in-place'' version of Stochastic Gradient Descent (IP-SGD) and Memory-Efficient Zeroth-order Optimizer (MeZO) have been proposed as solutions to improve memory efficiency. However, IP-SGD still requires a decent amount of memory, and MeZO suffers from slow convergence and degraded final performance due to its zeroth-order nature. This paper introduces Addax, a novel method that improves both memory efficiency and algorithm performance of IP-SGD by integrating it with MeZO. Specifically, Addax computes the zeroth-order or first-order gradient of the data points in the minibatch based on their memory consumption and combines zeroth- and first-order gradient estimates to obtain the updated direction in each step.By computing the zeroth-order order gradient of data points that require more memory and the first-order gradient of the ones that require less memory, Addax overcomes the slow convergence of MeZO and excessive memory requirement of IP-SGD. Additionally, the zeroth-order gradient acts as a regularizer for the first-order gradient, further enhancing the model's final performance.Theoretically, we establish the convergence of Addax under mild assumptions, demonstrating faster convergence and less restrictive hyper-parameter choices than MeZO. Our extensive experiments with diverse LMs and tasks show that Addax consistently outperforms MeZO in terms of accuracy and convergence speed, while having a comparable memory footprint. In particular, our experiments using one A100 GPU on OPT-13B model reveal that, on average, Addax outperforms MeZO in terms of accuracy/F1 score by 14%, and runs $15\times$ faster, while having a comparable memory footprint to MeZO. In our experiments on the larger OPT-30B model, on average, Addax outperforms MeZO in terms of accuracy/F1 score by >16% and runs $30\times$ faster on a single H100 GPU. Moreover, Addax surpasses the performance of standard fine-tuning approaches, such as IP-SGD and Adam, in most tasks in terms of Accuracy/F1 score with significantly less memory requirement.

## 2485. On Minimizing Adversarial Counterfactual Error in Adversarial Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28925 abstract： Deep Reinforcement Learning (DRL) policies are highly susceptible to adversarial noise in observations, which poses significant risks in safety-critical scenarios. The challenge inherent to adversarial perturbations is that by altering the information observed by the agent, the state becomes only partially observable. Existing approaches address this by either enforcing consistent actions across nearby states or maximizing the worst-case value within adversarially perturbed observations. However, the former suffers from performance degradation when attacks succeed, while the latter tends to be overly conservative, leading to suboptimal performance in benign settings. We hypothesize that these limitations stem from their failing to account for partial observability directly. To this end, we introduce a novel objective called Adversarial Counterfactual Error (ACoE), defined on the beliefs about the true state and balancing value optimization with robustness. To make ACoE scalable in model-free settings, we propose the theoretically-grounded surrogate objective Cumulative-ACoE (C-ACoE). Our empirical evaluations on standard benchmarks (MuJoCo, Atari, and Highway) demonstrate that our method significantly outperforms current state-of-the-art approaches for addressing adversarial RL challenges, offering

a promising direction for improving robustness in DRL under adversarial conditions. Our code is available at https://github.com/romanbelaire/acoe-robust-rl.

## 2486. Revisiting Nearest Neighbor for Tabular Data: A Deep Tabular Baseline Two Decades Later

链接：https://iclr.cc/virtual/2025/poster/30081 abstract： The widespread enthusiasm for deep learning has recently expanded into the domain of tabular data. Recognizing that the advancement in deep tabular methods is often inspired by classical methods, e.g., integration of nearest neighbors into neural networks, we investigate whether these classical methods can be revitalized with modern techniques.We revisit a differentiable version of $K$-nearest neighbors (KNN) --- Neighbourhood Components Analysis (NCA) --- originally designed to learn a linear projection to capture semantic similarities between instances, and seek to gradually add modern deep learning techniques on top. Surprisingly, our implementation of NCA using SGD and without dimensionality reduction already achieves decent performance on tabular data, in contrast to the results of using existing toolboxes like scikit-learn. Further equipping NCA with deep representations and additional training stochasticity significantly enhances its capability, being on par with the leading tree-based method CatBoost and outperforming existing deep tabular models in both classification and regression tasks on 300 datasets. We conclude our paper by analyzing the factors behind these improvements, including loss functions, prediction strategies, and deep architectures.The code is available at https://github.com/LAMDA-Tabular/TALENT.

## 2487. Multimodality Helps Few-shot 3D Point Cloud Semantic Segmentation

链接：https://iclr.cc/virtual/2025/poster/28634 abstract： Few-shot 3D point cloud segmentation (FS-PCS) aims at generalizing models to segment novel categories with minimal annotated support samples. While existing FS-PCS methods have shown promise, they primarily focus on unimodal point cloud inputs, overlooking the potential benefits of leveraging multimodal information. In this paper, we address this gap by introducing a multimodal FS-PCS setup, utilizing textual labels and the potentially available 2D image modality. Under this easy-to-achieve setup, we present the MultiModal Few-Shot SegNet (MM-FSS), a model effectively harnessing complementary information from multiple modalities. MM-FSS employs a shared backbone with two heads to extract intermodal and unimodal visual features, and a pretrained text encoder to generate text embeddings. To fully exploit the multimodal information, we propose a Multimodal Correlation Fusion (MCF) module to generate multimodal correlations, and a Multimodal Semantic Fusion (MSF) module to refine the correlations using text-aware semantic guidance. Additionally, we propose a simple yet effective Test-time Adaptive Cross-modal Calibration (TACC) technique to mitigate training bias, further improving generalization. Experimental results on S3DIS and ScanNet datasets demonstrate significant performance improvements achieved by our method. The efficacy of our approach indicates the benefits of leveraging commonly-ignored free modalities for FS-PCS, providing valuable insights for future research. The code is available at github.com/ZhaochongAn/Multimodality-3D-Few-Shot.

## 2488. Limits to scalable evaluation at the frontier: LLM as judge won’t beat twice the data

链接：https://iclr.cc/virtual/2025/poster/29881 abstract： High quality annotations are increasingly a bottleneck in the explosively growing machine learning ecosystem. Scalable evaluation methods that avoid costly annotation have therefore become an important research ambition. Many hope to use strong existing models in lieu of costly labels to provide cheap model evaluations. Unfortunately, this method of using models as judges introduces biases, such as self-preferencing, that can distort model comparisons. An emerging family of debiasing tools promises to fix these issues by using a few high quality labels to debias a large number of model judgments. In this paper, we study how far such debiasing methods, in principle, can go. Our main result shows that when the judge is no more accurate than the evaluated model, no debiasing method can decrease the required amount of ground truth labels by more than half. Our result speaks to the severe limitations of the LLM-as-a-judge paradigm at the evaluation frontier where the goal is to assess newly released models that are possibly better than the judge. Through an empirical evaluation, we demonstrate that the sample size savings achievable in practice are even more modest than what our theoretical limit suggests. Along the way, our work provides new observations about debiasing methods for model evaluation, and points out promising avenues for future work.

## 2489. Parameter and Memory Efficient Pretraining via Low-rank Riemannian Optimization

链接：https://iclr.cc/virtual/2025/poster/28725 abstract： Pretraining large language models often requires significant computational resources and memory due to their vast parameter amount. An effective approach to enhance parameter efficiency in both training and inference is to parameterize each full-size weight as the product of two trainable low-rank factors. While low-rank fine-tuning has achieved great success, low-rank pretraining remains challenging as it requires learning extensive knowledge from scratch under the restrictive low-rank parameterization. During standard low-rank pretraining, separately optimizing the low-rank factors introduces redundant information from the full gradient, which hinders the learning process. To achieve efficient yet effective low-rank pretraining, we propose a **Lo**w-rank **R**iemannian **O**ptimizer (**LORO**). At each LORO update step, the low-rank factor pairs are jointly updated to ensure their full-size product moves along the steepest descent direction on the low-rank manifold, without the need to compute any memory-intensive full-size matrices or gradients.

Hence, our LORO finds low-rank models that achieve high performance comparable to full-size pretrained models, while significantly reducing memory usage and accelerating both training and inference. A LLaMA 1B model pretrained with LORO achieves a perplexity score of 2\% better than the full-size baseline, with a 54\% reduction in model memory, a $\times1.8$ speedup in training, and a $\times2.2$ speedup in inference. The code is available on https://github.com/mzf666/LORO-main.

## 2490. Reward Learning from Multiple Feedback Types

链接：https://iclr.cc/virtual/2025/poster/30703 abstract： Learning rewards from preference feedback has become an important tool in the alignment of agentic models. Preference-based feedback, often implemented as a binary comparison between multiple completions, is an established method to acquire large-scale human feedback. However, human feedback in other contexts is often much more diverse. Such diverse feedback can better support the goals of a human annotator, and the simultaneous use of multiple sources might be mutually informative for the learning process or carry type-dependent biases for the reward learning process.Despite these potential benefits, learning from different feedback types has yet to be explored extensively.In this paper, we bridge this gap by enabling experimentation and evaluating multi-type feedback in a wide set of environments. We present a process to generate high-quality simulated feedback of six different types. Then, we implement reward models and downstream RL training for all six feedback types.Based on the simulated feedback, we investigate the use of types of feedback across ten RL environments and compare them to pure preference-based baselines. We show empirically that diverse types of feedback can be utilized and lead to strong reward modeling performance. This work is the first strong indicator of the potential of multi-type feedback for RLHF.

## 2491. Probabilistic Neural Pruning via Sparsity Evolutionary Fokker-Planck-Kolmogorov Equation

链接：https://iclr.cc/virtual/2025/poster/28772 abstract： Neural pruning aims to compress and accelerate deep neural networks by identifying the optimal subnetwork within a specified sparsity budget. In this work, we study how to gradually sparsify the unpruned dense model to the target sparsity level with minimal performance drop. Specifically, we analyze the evolution of the population of optimal subnetworks under continuous sparsity increments from a thermodynamic perspective. We first reformulate neural pruning as an expected loss minimization problem over the mask distributions. Then, we establish an effective approximation for the sparsity evolution of the optimal mask distribution, termed the Sparsity Evolutionary Fokker-Planck-Kolmogorov Equation (SFPK), which provides closed-form, mathematically tractable guidance on distributional transitions for minimizing the expected loss under an infinitesimal sparsity increment. On top of that, we propose SFPK-pruner, a particle simulation-based probabilistic pruning method, to sample performant masks with desired sparsity from the destination distribution of SFPK. In theory, we establish the convergence guarantee for the proposed SFPK-pruner. Our SFPK-pruner exhibits competitive performance in various pruning scenarios. The code is available on https://github.com/mzf666/SFPK-main.

## 2492. Training on the Test Task Confounds Evaluation and Emergence

链接：https://iclr.cc/virtual/2025/poster/28639 abstract： We study a fundamental problem in the evaluation of large language models that we call training on the test task. Unlike wrongful practices like training on the test data, leakage, or data contamination, training on the test task is not a malpractice. Rather, the term describes a growing set of techniques to include task-relevant data in the pretraining stage of a language model. We demonstrate that training on the test task confounds both relative model evaluations and claims about emergent capabilities. We argue that the seeming superiority of one model family over another may be explained by a different degree of training on the test task. To this end, we propose an effective method to adjust for the effect of training on the test task on benchmark evaluations. Put simply, to fine-tune each model under comparison on the same task-relevant data before evaluation. Lastly, we show that instances of emergent behavior disappear gradually as models train on the test task. Our work promotes a new perspective on the evaluation of large language models with broad implications for benchmarking and the study of emergent capabilities.

## 2493. Homomorphism Expressivity of Spectral Invariant Graph Neural Networks

链接：https://iclr.cc/virtual/2025/poster/28183 abstract： Graph spectra are an important class of structural features on graphs that have shown promising results in enhancing Graph Neural Networks (GNNs). Despite their widespread practical use, the theoretical understanding of the power of spectral invariants --- particularly their contribution to GNNs --- remains incomplete. In this paper, we address this fundamental question through the lens of homomorphism expressivity, providing a comprehensive and quantitative analysis of the expressive power of spectral invariants. Specifically, we prove that spectral invariant GNNs can homomorphism-count exactly a class of specific tree-like graphs which we refer to as \emph{parallel trees}. We highlight the significance of this result in various contexts, including establishing a quantitative expressiveness hierarchy across different architectural variants, offering insights into the impact of GNN depth, and understanding the subgraph counting capabilities of spectral invariant GNNs. In particular, our results significantly extend \citet{arvind2024hierarchy} and settle their open questions. Finally, we generalize our analysis to higher-order GNNs and answer an open question raised by \citet{zhang2024expressive}.

## 2494. Towards Marginal Fairness Sliced Wasserstein Barycenter

链接：https://iclr.cc/virtual/2025/poster/29877 abstract： The Sliced Wasserstein barycenter (SWB) is a widely acknowledged method for efficiently generalizing the averaging operation within probability measure spaces. However, achieving marginal fairness SWB, ensuring approximately equal distances from the barycenter to marginals, remains unexplored. The uniform weighted SWB is not necessarily the optimal choice to obtain the desired marginal fairness barycenter due to the heterogeneous structure of marginals and the non-optimality of the optimization. As the first attempt to tackle the problem, we define the marginal fairness sliced Wasserstein barycenter (MFSWB) as a constrained SWB problem. Due to the computational disadvantages of the formal definition, we propose two hyperparameter-free and computationally tractable surrogate MFSWB problems that implicitly minimize the distances to marginals and encourage marginal fairness at the same time. To further improve the efficiency, we perform slicing distribution selection and obtain the third surrogate definition by introducing a new slicing distribution that focuses more on marginally unfair projecting directions. We discuss the relationship of the three proposed problems and their relationship to sliced multi-marginal Wasserstein distance. Finally, we conduct experiments on finding 3D point-clouds averaging, color harmonization, and training of sliced Wasserstein autoencoder with class-fairness representation to show the favorable performance of the proposed surrogate MFSWB problems.

## 2495. How Do Large Language Models Understand Graph Patterns? A Benchmark for Graph Pattern Comprehension

链接：https://iclr.cc/virtual/2025/poster/30499 abstract： Benchmarking the capabilities and limitations of large language models (LLMs) in graph-related tasks is becoming an increasingly popular and crucial area of research. Recent studies have shown that LLMs exhibit a preliminary ability to understand graph structures and node features. However, the potential of LLMs in graph pattern mining remains largely unexplored. This is a key component in fields such as computational chemistry, biology, and social network analysis. To bridge this gap, this work introduces a comprehensive benchmark to assess LLMs' capabilities in graph pattern tasks. We have developed a benchmark that evaluates whether LLMs can understand graph patterns based on either terminological or topological descriptions. Additionally, our benchmark tests the LLMs' capacity to autonomously discover graph patterns from data. The benchmark encompasses both synthetic and real datasets, and a variety of models, with a total of 11 tasks and 7 models. Our experimental framework is designed for easy expansion to accommodate new models and datasets. Our findings reveal that: (1) LLMs have preliminary abilities to understand graph patterns, with O1-mini outperforming in the majority of tasks; (2) Formatting input graph data to align with the knowledge acquired during pretraining can enhance performance; (3) LLMs employ diversepotential algorithms to solve one task, with performance varying based on their execution capabilities.

## 2496. MovieDreamer: Hierarchical Generation for Coherent Long Visual Sequences

链接：https://iclr.cc/virtual/2025/poster/29698 abstract： Recent advancements in video generation have primarily leveraged diffusion models for short-duration content. However, these approaches often fall short in modeling complex narratives and maintaining character consistency over extended periods, which is essential for long-form video production like movies. We propose MovieDreamer, a novel hierarchical framework that integrates the strengths of autoregressive models with diffusion-based rendering to pioneer long-duration video generation with intricate plot progressions and high visual fidelity. Our approach utilizes autoregressive models for global narrative coherence, predicting sequences of visual tokens that are subsequently transformed into high-quality video frames through diffusion rendering. This method is akin to traditional movie production processes, where complex stories are factorized down into manageable scene capturing. Further, we employ a multimodal script that enriches scene descriptions with detailed character information and visual style, enhancing continuity and character identity across scenes. We present extensive experiments across various movie genres, demonstrating that our approach not only achieves superior visual and narrative quality but also effectively extends the duration of generated content significantly beyond current capabilities.

## 2497. CO-MOT: Boosting End-to-end Transformer-based Multi-Object Tracking via Coopetition Label Assignment and Shadow Sets

链接：https://iclr.cc/virtual/2025/poster/31235 abstract： Existing end-to-end Multi-Object Tracking (e2e-MOT) methods have not surpassed non-end-to-end tracking-by-detection methods. One possible reason lies in the training label assignment strategy that consistently binds the tracked objects with tracking queries and assigns few newborns to detection queries. Such an assignment, with one-to-one bipartite matching, yields an unbalanced training, i.e., scarce positive samples for detection queries, especially for an enclosed scene with the majority of the newborns at the beginning of videos. As such, e2e-MOT will incline to generate a tracking terminal without renewal or re-initialization, compared to other tracking-by-detection methods.To alleviate this problem, we propose Co-MOT, a simple yet effective method to facilitate e2e-MOT by a novel coopetition label assignment with a shadow concept. Specifically, we add tracked objects to the matching targets for detection queries when performing the label assignment for training the intermediate decoders. For query initialization, we expand each query by a set of shadow counterparts with limited disturbance to itself.With extensive ablation studies, Co-MOT achieves superior performances without extra costs, e.g., 69.4% HOTA on DanceTrack and 52.8% TETA on BDD100K. Impressively, Co-MOT only requires 38% FLOPs of MOTRv2 with comparable performances, resulting in the 1.4× faster inference speed. Source code is publicly available at GitHub.

## 2498. Does Spatial Cognition Emerge in Frontier Models?

链接：https://iclr.cc/virtual/2025/poster/29371 abstract： Not yet. We present SPACE, a benchmark that systematically evaluates spatial cognition in frontier models. Our benchmark builds on decades of research in cognitive science. It evaluates large-scale mapping abilities that are brought to bear when an organism traverses physical environments, smaller-scale reasoning about object shapes and layouts, and cognitive infrastructure such as spatial attention and memory. For many tasks, we instantiate parallel presentations via text and images, allowing us to benchmark both large language models and large multimodal models. Results suggest that contemporary frontier models fall short of the spatial intelligence of animals, performing near chance level on a number of classic tests of animal cognition.

## 2499. Chemistry-Inspired Diffusion with Non-Differentiable Guidance

链接：https://iclr.cc/virtual/2025/poster/31002 abstract： Recent advances in diffusion models have shown remarkable potential in the conditional generation of novel molecules. These models can be guided in two ways: (i) explicitly, through additional features representing the condition, or (ii) implicitly, using a property predictor. However, training property predictors or conditional diffusion models requires an abundance of labeled data and is inherently challenging in real-world applications. We propose a novel approach that attenuates the limitations of acquiring large labeled datasets by leveraging domain knowledge from quantum chemistry as a non-differentiable oracle to guide an unconditional diffusion model. Instead of relying on neural networks, the oracle provides accurate guidance in the form of estimated gradients, allowing the diffusion process to sample from a conditional distribution specified by quantum chemistry. We show that this results in more precise conditional generation of novel and stable molecular structures. Our experiments demonstrate that our method: (1) significantly reduces atomic forces, enhancing the validity of generated molecules when used for stability optimization; (2) is compatible with both explicit and implicit guidance in diffusion models, enabling joint optimization of molecular properties and stability; and (3) generalizes effectively to molecular optimization tasks beyond stability optimization. Our implementation is available at https://github.com/A-Chicharito-S/ChemGuide.

## 2500. Cut Your Losses in Large-Vocabulary Language Models

链接：https://iclr.cc/virtual/2025/poster/30424 abstract： As language models grow ever larger, so do their vocabularies.This has shifted the memory footprint of LLMs during training disproportionately to one single layer: the cross-entropy in the loss computation.Cross-entropy builds up a logit matrix with entries for each pair of input tokens and vocabulary items and, for small models, consumes an order of magnitude more memory than the rest of the LLM combined.We propose Cut Cross-Entropy (CCE), a method that computes the cross-entropy loss without materializing the logits for all tokens into global memory.Rather, CCE only computes the logit for the correct token and evaluates the log-sum-exp over all logits on the fly.We implement a custom kernel that performs the matrix multiplications and the log-sum-exp reduction over the vocabulary in flash memory, making global memory consumption for the cross-entropy computation negligible. This has a dramatic effect. Taking the Gemma 2 (2B) model as an example, CCE reduces the memory footprint of the loss computation from 24 GB to 1 MB, and the total training-time memory consumption of the classifier head from 28 GB to 1 GB.To improve the throughput of CCE, we leverage the inherent sparsity of softmax and propose to skip elements of the gradient computation that have a negligible (i.e. below numerical precision) contribution to the gradient.Experiments demonstrate that the dramatic reduction in memory consumption is accomplished without sacrificing training speed or convergence.

## 2501. COME: Test-time Adaption by Conservatively Minimizing Entropy

链接：https://iclr.cc/virtual/2025/poster/30977 abstract： Machine learning models must continuously self-adjust themselves for novel data distribution in the open world. As the predominant principle, entropy minimization (EM) has been proven to be a simple yet effective cornerstone in existing test-time adaption (TTA) methods. While unfortunately its fatal limitation (i.e., overconfidence) tends to result in model collapse. For this issue, we propose to \textbf{\texttt{Co}}nservatively \textbf{\texttt{M}}inimize the \textbf{\texttt{E}}ntropy (\texttt{COME}), which is a simple drop-in replacement of traditional EM to elegantly address the limitation. In essence, \texttt{COME} explicitly models the uncertainty by characterizing a Dirichlet prior distribution over model predictions during TTA. By doing so, \texttt{COME} naturally regularizes the model to favor conservative confidence on unreliable samples. Theoretically, we provide a preliminary analysis to reveal the ability of \texttt{COME} in enhancing the optimization stability by introducing a data-adaptive lower bound on the entropy. Empirically, our method achieves state-of-the-art performance on commonly used benchmarks, showing significant improvements in terms of classification accuracy and uncertainty estimation under various settings including standard, life-long and open-world TTA, i.e., up to $34.5\%$ improvement on accuracy and $15.1\%$ on false positive rate. Our code is available at: \href{https://github.com/BlueWhaleLab/COME}{https://github.com/BlueWhaleLab/COME}.

## 2502. CoMotion: Concurrent Multi-person 3D Motion

链接：https://iclr.cc/virtual/2025/poster/28254 abstract： We introduce an approach for detecting and tracking detailed 3D poses of multiple people from a single monocular camera stream. Our system maintains temporally coherent predictions in crowded scenes filled with difficult poses and occlusions. Our model performs both strong per-frame detection and a learned pose update to track people from frame to frame. Rather than match detections across time, poses are updated directly from a new input image, which enables online tracking through occlusion. We train on numerous image and video datasets leveraging pseudo-labeled annotations to produce a model that matches state-of-the-art systems in 3D pose estimation accuracy while

being faster and more accurate in tracking multiple people through time.

## 2503. GNNs Getting ComFy: Community and Feature Similarity Guided Rewiring

链接：https://iclr.cc/virtual/2025/poster/28833 abstract： Maximizing the spectral gap through graph rewiring has been proposed to enhance the performance of message-passing graph neural networks (GNNs) by addressing over-squashing. However, as we show, minimizing the spectral gap can also improve generalization. To explain this, we analyze how rewiring can benefit GNNs within the context of stochastic block models. Since spectral gap optimization primarily influences community strength, it improves performance when the community structure aligns with node labels. Building on this insight, we propose three distinct rewiring strategies that explicitly target community structure, node labels, and their alignment: (a) community structure-based rewiring (ComMa), a more computationally efficient alternative to spectral gap optimization that achieves similar goals; (b) feature similarity-based rewiring (FeaSt), which focuses on maximizing global homophily; and (c) a hybrid approach (ComFy), which enhances local feature similarity while preserving community structure to optimize label-community alignment. Extensive experiments confirm the effectiveness of these strategies and support our theoretical insights.

## 2504. MMRole: A Comprehensive Framework for Developing and Evaluating Multimodal Role-Playing Agents

链接：https://iclr.cc/virtual/2025/poster/30350 abstract： Recently, Role-Playing Agents (RPAs) have garnered increasing attention for their potential to deliver emotional value and facilitate sociological research.However, existing studies are primarily confined to the textual modality, unable to simulate humans' multimodal perceptual capabilities.To bridge this gap, we introduce the concept of Multimodal Role-Playing Agents (MRPAs), and propose a comprehensive framework, MMRole, for their development and evaluation, which comprises a personalized multimodal dataset and a robust evaluation approach.Specifically, we construct a large-scale, high-quality dataset, MMRole-Data, consisting of 85 characters, 11K images, and 14K single or multi-turn dialogues.Additionally, we present a robust evaluation approach, MMRole-Eval, encompassing eight metrics across three dimensions, where a reward model is designed to score MRPAs with the constructed ground-truth data for comparison.Moreover, we develop the first specialized MRPA, MMRole-Agent.Extensive evaluation results demonstrate the improved performance of MMRole-Agent and highlight the primary challenges in developing MRPAs, emphasizing the need for enhanced multimodal understanding and role-playing consistency.The data, code, and models are all available at https://github.com/YanqiDai/MMRole.

## 2505. DRoC: Elevating Large Language Models for Complex Vehicle Routing via Decomposed Retrieval of Constraints

链接：https://iclr.cc/virtual/2025/poster/28142 abstract： This paper proposes Decomposed Retrieval of Constraints (DRoC), a novel framework aimed at enhancing large language models (LLMs) in exploiting solvers to tackle vehicle routing problems (VRPs) with intricate constraints. While LLMs have shown promise in solving simple VRPs, their potential in addressing complex VRP variants is still suppressed, due to the limited embedded internal knowledge that is required to accurately reflect diverse VRP constraints. Our approach mitigates the issue by integrating external knowledge via a novel retrieval-augmented generation (RAG) approach. More specifically, the DRoC decomposes VRP constraints, externally retrieves information relevant to each constraint, and synergistically combines internal and external knowledge to benefit the program generation for solving VRPs. The DRoC also allows LLMs to dynamically select between RAG and self-debugging mechanisms, thereby optimizing program generation without the need for additional training. Experiments across 48 VRP variants exhibit the superiority of DRoC, with significant improvements in the accuracy rate and runtime error rate delivered by the generated programs. The DRoC framework has the potential to elevate LLM performance in complex optimization tasks, fostering the applicability of LLMs in industries such as transportation and logistics.

## 2506. Oracle efficient truncated statistics

链接：https://iclr.cc/virtual/2025/poster/29220 abstract： We study the problem of learning from truncated samples: instead of observingsamples from some underlying population $p^\ast$, we observe only the examples that fall in some survival set $S \subset \mathbb{R}^d$ whose probability mass (measured with respect to $p^\ast$) is at least $\alpha$. Assuming membership oracle access to the truncation set $S$, prior works obtained algorithms for the case where $p^\ast$ is Gaussian or more generally an exponential family with strongly convex likelihood --- albeit with a super-polynomial dependency on the (inverse) survival mass $1/\alpha$both in terms of runtime and in number of oracle calls to the set $S$. In this work we design a new learning method with runtime and query complexity polynomial in $1/\alpha$. Our result significantly improves over the prior works by focusing on efficiently solving the underlying optimization problem using a generalpurpose optimization algorithm with minimal assumptions.

## 2507. Training Free Guided Flow-Matching with Optimal Control

链接：https://iclr.cc/virtual/2025/poster/30912 abstract： Controlled generation with pre-trained Diffusion and Flow Matching models has vast applications. One strategy for guiding ODE-based generative models is through optimizing a target loss

$R(x_1)$ while staying close to the prior distribution. Along this line, some recent work showed the effectiveness of guiding flow model by differentiating through its ODE sampling process. Despite the superior performance, the theoretical understanding of this line of methods is still preliminary, leaving space for algorithm improvement. Moreover, existing methods predominately focus on Euclidean data manifold, and there is a compelling need for guided flow methods on complex geometries such as SO(3), which prevails in high-stake scientific applications like protein design. We present OC-Flow, a general and theoretically grounded training-free framework for guided flow matching using optimal control. Building upon advances in optimal control theory, we develop effective and practical algorithms for solving optimal control in guided ODE-based generation and provide a systematic theoretical analysis of the convergence guarantee in both Euclidean and SO(3). We show that existing backprop-through-ODE methods can be interpreted as special cases of Euclidean OC-Flow. OC-Flow achieved superior performance in extensive experiments on text-guided image manipulation, conditional molecule generation, and all-atom peptide design.

## 2508. 6D Object Pose Tracking in Internet Videos for Robotic Manipulation

链接：https://iclr.cc/virtual/2025/poster/31220 abstract： We seek to extract a temporally consistent 6D pose trajectory of a manipulated object from an Internet instructional video. This is a challenging set-up for current 6D pose estimation methods due to uncontrolled capturing conditions, subtle but dynamic object motions, and the fact that the exact mesh of the manipulated object is not known. To address these challenges, we present the following contributions. First, we develop a new method that estimates the 6D pose of any object in the input image without prior knowledge of the object itself. The method proceeds by (i) retrieving a CAD model similar to the depicted object from a large-scale model database, (ii) 6D aligning the retrieved CAD model with the input image, and (iii) grounding the absolute scale of the object with respect to the scene. Second, we extract smooth 6D object trajectories from Internet videos by carefully tracking the detected objects across video frames. The extracted object trajectories are then retargeted via trajectory optimization into the configuration space of a robotic manipulator. Third, we thoroughly evaluate and ablate our 6D pose estimation method on YCB-V and HOPE-Video datasets as well as a new dataset of instructional videos manually annotated with approximate 6D object trajectories. We demonstrate significant improvements over existing state-of-the-art RGB 6D pose estimation methods. Finally, we show that the 6D object motion estimated from Internet videos can be transferred to a 7-axis robotic manipulator both in a virtual simulator as well as in a real world set-up. We also successfully apply our method to egocentric videos taken from the EPIC-KITCHENS dataset, demonstrating potential for Embodied AI applications.

## 2509. SaLoRA: Safety-Alignment Preserved Low-Rank Adaptation

链接：https://iclr.cc/virtual/2025/poster/30283 abstract： As advancements in large language models (LLMs) continue and the demand for personalized models increases, parameter-efficient fine-tuning (PEFT) methods (e.g., LoRA) become essential due to their efficiency in reducing computation costs.However, recent studies have raised alarming concerns that LoRA fine-tuning could potentially compromise the safety alignment in LLMs, posing significant risks for the model owner.In this paper, we first investigate the underlying mechanism by analyzing the changes in safety alignment related features before and after fine-tuning.Then, we propose a fixed safety module calculated by safety data and a task-specific initialization for trainable parameters in low-rank adaptations, termed Safety-alignment preserved Low-Rank Adaptation (SaLoRA). Unlike previous LoRA methods and their variants, SaLoRA enables targeted modifications to LLMs without disrupting their original alignments. Our experiments show that SaLoRA outperforms various adapters-based approaches across various evaluation metrics in different fine-tuning tasks.

## 2510. MGCFNN: A Neural MultiGrid Solver with Novel Fourier Neural Network for High Wave Number Helmholtz Equations

链接：https://iclr.cc/virtual/2025/poster/29525 abstract： Solving high wavenumber Helmholtz equations is notoriously challenging. Traditional solvers have yet to yield satisfactory results, and most neural network methods struggle to accurately solve cases with extremely high wavenumbers within heterogeneous media. This paper presents an advanced multigrid-hierarchical AI solver, tailored specifically for high wavenumber Helmholtz equations. We adapt the MGCNN architecture to align with the problem setting and incorporate a novel Fourier neural network (FNN) to match the characteristics of Helmholtz equations. FNN, mathematically akin to the convolutional neural network (CNN), enables faster propagation of source influence during the solve phase, making it particularly suitable for handling large size, high wavenumber problems. We conduct supervised learning tests against numerous neural operator learning methods to demonstrate the superior learning capabilities of our solvers. Additionally, we perform scalability tests using an unsupervised strategy to highlight our solvers' significant speedup over the most recent specialized AI solver and AI-enhanced traditional solver for high wavenumber Helmholtz equations. We also carry out an ablation study to underscore the effectiveness of the multigrid hierarchy and the benefits of introducing FNN. Notably, our solvers exhibit optimal convergence of $\mathcal{O}(k)$ up to $k \approx 2000$.

## 2511. BTBS-LNS: Binarized-Tightening, Branch and Search on Learning LNS Policies for MIP

链接：https://iclr.cc/virtual/2025/poster/28102 abstract： Learning to solve large-scale Mixed Integer Program (MIP) problems is an emerging research topic, and policy learning-based Large Neighborhood Search (LNS) has been a popular paradigm. However, the explored space of LNS policy is often limited even in the training phase, making the learned policy sometimes wrongly fix some potentially important variables early in the search, leading to local optimum in some cases. Moreover, many

methods only assume binary variables to deal with. We present a practical approach, termed Binarized-Tightening Branch-and-Search for Large Neighborhood Search (BTBS-LNS). It comprises three key techniques: 1) the ``Binarized Tightening'' technique for integer variables to handle their wide range by binary encoding and bound tightening; 2) an attention-based tripartite graph to capture global correlations among variables and constraints for an MIP instance; 3) an extra branching network as a global view, to identify and optimize wrongly-fixed backdoor variables at each search step. Experiments show its superior performance over the open-source solver SCIP and LNS baselines. Moreover, it performs competitively with, and sometimes better than the commercial solver Gurobi (v9.5.0), especially on the MIPLIB2017 benchmark chosen by Hans Mittelmann, where our method can deliver 10\% better primal gaps compared with Gurobi in a 300s cut-off time.

## 2512. How new data permeates LLM knowledge and how to dilute it

链接：https://iclr.cc/virtual/2025/poster/29889 abstract： Large language models continually learn through the accumulation of gradient-based updates, but how individual pieces of new information affect existing knowledge, leading to both beneficial generalization and problematic hallucination, remains poorly understood. We demonstrate that when learning new information, LLMs exhibit a "priming" effect: learning a new fact can cause the model to inappropriately apply that knowledge in unrelated contexts.To systematically study this phenomenon, we introduce "Outlandish," a carefully curated dataset of 1320 diverse text samples designed to probe how new knowledge permeates through an LLM's existing knowledge base. Using this dataset, we show that the degree of priming after learning new information can be predicted by measuring the token probability of key words before training. This relationship holds robustly across different model architectures (PALM-2, Gemma, Llama), sizes, and training stages.Finally, we develop two novel techniques to modulate how new knowledge affects existing model behavior: (1) a stepping-stone" text augmentation strategy and (2) anignore-k" update pruning method. These approaches reduce undesirable priming effects by 50-95% while preserving the model's ability to learn new information. Our findings provide both empirical insights into how LLMs learn and practical tools for improving the specificity of knowledge insertion in language models. Further materials: https://sunchipsster1.github.io/projects/outlandish/

## 2513. Pre-training of Foundation Adapters for LLM Fine-tuning

链接：https://iclr.cc/virtual/2025/poster/31361 abstract： Adapter-based fine-tuning methods insert small, trainable adapters into frozen pre-trained LLMs, significantly reducing computational costs while maintaining performance. However, despite these advantages, traditional adapter fine-tuning suffers from training instability due to random weight initialization. This instability can lead to inconsistent performance across different runs. Therefore, to address this issue, this blog post introduces pre-trained foundation adapters as a technique for weight initialization. This technique potentially improves the efficiency and effectiveness of the fine-tuning process. Specifically, we combine continual pre-training and knowledge distillation to pre-train foundation adapters. Experiments confirm the effectiveness of this approach across multiple tasks. Moreover, we highlight the advantage of using pre-trained foundation adapter weights over random initialization specifically in a summarization task.

## 2514. Spiking Vision Transformer with Saccadic Attention

链接：https://iclr.cc/virtual/2025/poster/28227 abstract： The combination of Spiking Neural Networks (SNNs) and Vision Transformers (ViTs) holds potential for achieving both energy efficiency and high performance, particularly suitable for edge vision applications. However, a significant performance gap still exists between SNN-based ViTs and their ANN counterparts. Here, we first analyze why SNN-based ViTs suffer from limited performance and identify a mismatch between the vanilla self-attention mechanism and spatio-temporal spike trains. This mismatch results in degraded spatial relevance and limited temporal interactions. To address these issues, we draw inspiration from biological saccadic attention mechanisms and introduce an innovative Saccadic Spike Self-Attention (SSSA) method. Specifically, in the spatial domain, SSSA employs a novel spike distribution-based method to effectively assess the relevance between Query and Key pairs in SNN-based ViTs. Temporally, SSSA employs a saccadic interaction module that dynamically focuses on selected visual areas at each timestep and significantly enhances whole scene understanding through temporal interactions.Building on the SSSA mechanism, we develop a SNN-based Vision Transformer (SNN-ViT). Extensive experiments across various visual tasks demonstrate that SNN-ViT achieves state-of-the-art performance with linear computational complexity. The effectiveness and efficiency of the SNN-ViT highlight its potential for power-critical edge vision applications.

## 2515. NExT-Mol: 3D Diffusion Meets 1D Language Modeling for 3D Molecule Generation

链接：https://iclr.cc/virtual/2025/poster/28319 abstract： 3D molecule generation is crucial for drug discovery and material design. While prior efforts focus on 3D diffusion models for their benefits in modeling continuous 3D conformers, they overlook the advantages of 1D SELFIES-based Language Models (LMs), which can generate 100\% valid molecules and leverage the billion-scale 1D molecule datasets. To combine these advantages for 3D molecule generation, we propose a foundation model -- NExT-Mol: 3D Diffusion Meets 1D Language Modeling for 3D Molecule Generation. NExT-Mol uses an extensively pretrained molecule LM for 1D molecule generation, and subsequently predicts the generated molecule's 3D conformers with a 3D diffusion model. We enhance NExT-Mol's performance by scaling up the LM's model size, refining the diffusion neural architecture, and applying 1D to 3D transfer learning. Notably, our 1D molecule LM significantly outperforms baselines in distributional similarity while ensuring validity, and our 3D diffusion model achieves leading performances in conformer prediction. Given these improvements in 1D and 3D modeling, NExT-Mol achieves a 26\% relative improvement in 3D FCD for de novo 3D generation on GEOM-DRUGS, and a 13\% average relative gain for conditional 3D generation on QM9-2014. Our

codes and pretrained checkpoints are available at https://github.com/acharkq/NExT-Mol.

## 2516. NeuroLM: A Universal Multi-task Foundation Model for Bridging the Gap between Language and EEG Signals

链接：https://iclr.cc/virtual/2025/poster/30149 abstract： Recent advancements for large-scale pre-training with neural signals such as electroencephalogram (EEG) have shown promising results, significantly boosting the development of brain-computer interfaces (BCIs) and healthcare. However, these pre-trained models often require full fine-tuning on each downstream task to achieve substantial improvements, limiting their versatility and usability, and leading to considerable resource wastage. To tackle these challenges, we propose NeuroLM, the first multi-task foundation model that leverages the capabilities of Large Language Models (LLMs) by regarding EEG signals as a foreign language, endowing the model with multi-task learning and inference capabilities. Our approach begins with learning a text-aligned neural tokenizer through vector-quantized temporal-frequency prediction, which encodes EEG signals into discrete neural tokens. These EEG tokens, generated by the frozen vector-quantized (VQ) encoder, are then fed into an LLM that learns causal EEG information via multi-channel autoregression. Consequently, NeuroLM can understand both EEG and language modalities. Finally, multi-task instruction tuning adapts NeuroLM to various downstream tasks. We are the first to demonstrate that, by specific incorporation with LLMs, NeuroLM unifies diverse EEG tasks within a single model through instruction tuning. The largest variant NeuroLM-XL has record-breaking 1.7B parameters for EEG signal processing, and is pre-trained on a large-scale corpus comprising approximately 25,000-hour EEG data. When evaluated on six diverse downstream datasets, NeuroLM showcases the huge potential of this multi-task learning paradigm.

## 2517. Diffusion Feedback Helps CLIP See Better

链接：https://iclr.cc/virtual/2025/poster/28060 abstract： Contrastive Language-Image Pre-training (CLIP), which excels at abstracting open-world representations across domains and modalities, has become a foundation for a variety of vision and multimodal tasks. However, recent studies reveal that CLIP has severe visual shortcomings, such as which can hardly distinguish orientation, quantity, color, structure, etc. These visual shortcomings also limit the perception capabilities of multimodal large language models (MLLMs) built on CLIP. The main reason could be that the image-text pairs used to train CLIP are inherently biased, due to the lack of the distinctiveness of the text and the diversity of images. In this work, we present a simple post-training approach for CLIP models, which largely overcomes its visual shortcomings via a self-supervised diffusion process. We introduce DIVA, which uses the DIffusion model as a Visual Assistant for CLIP. Specifically, DIVA leverages generative feedback from text-to-image diffusion models to optimize CLIP representations, with only images (without corresponding text). We demonstrate that DIVA improves CLIP's performance on the challenging MMVP-VLM benchmark which assesses fine-grained visual abilities to a large extent (e.g., 3-7%), and enhances the performance of MLLMs and vision models on multimodal understanding and segmentation tasks. Extensive evaluation on 29 image classification and retrieval benchmarks confirms that our framework preserves CLIP's strong zero-shot capabilities. The code is publicly available at https://github.com/baaivision/DIVA.

## 2518. Stabilizing Reinforcement Learning in Differentiable Multiphysics Simulation

链接：https://iclr.cc/virtual/2025/poster/30460 abstract：

## 2519. Conformalized Interactive Imitation Learning: Handling Expert Shift and Intermittent Feedback

链接：https://iclr.cc/virtual/2025/poster/29247 abstract： In interactive imitation learning (IL), uncertainty quantification offers a way for the learner (i.e. robot) to contend with distribution shifts encountered during deployment by actively seeking additional feedback from an expert (i.e. human) online. Prior works use mechanisms like ensemble disagreement or Monte Carlo dropout to quantify when black-box IL policies are uncertain; however, these approaches can lead to overconfident estimates when faced with deployment-time distribution shifts. Instead, we contend that we need uncertainty quantification algorithms that can leverage the expert human feedback received during deployment time to adapt the robot's uncertainty online. To tackle this, we draw upon online conformal prediction, a distribution-free method for constructing prediction intervals online given a stream of ground-truth labels. Human labels, however, are intermittent in the interactive IL setting. Thus, from the conformal prediction side, we introduce a novel uncertainty quantification algorithm called intermittent quantile tracking (IQT) that leverages a probabilistic model of intermittent labels, maintains asymptotic coverage guarantees, and empirically achieves desired coverage levels. From the interactive IL side, we develop ConformalDAgger, a new approach wherein the robot uses prediction intervals calibrated by IQT as a reliable measure of deployment-time uncertainty to actively query for more expert feedback. We compare ConformalDAgger to prior uncertainty-aware DAgger methods in scenarios where the distribution shift is (and isn't) present because of changes in the expert's policy. We find that in simulated and hardware deployments on a 7DOF robotic manipulator, ConformalDAgger detects high uncertainty when the expert shifts and increases the number of interventions compared to baselines, allowing the robot to more quickly learn the new behavior.

## 2520. Agents' Room: Narrative Generation through Multi-step Collaboration

链接：https://iclr.cc/virtual/2025/poster/30214 abstract： Writing compelling fiction is a multifaceted process combining elements such as crafting a plot, developing interesting characters, and using evocative language. While large language models (LLMs) show promise for story writing, they currently rely heavily on intricate prompting, which limits their use. We propose Agents' Room, a generation framework inspired by narrative theory, that decomposes narrative writing into subtasks tackled by specialized agents. To illustrate our method, we introduce Tell Me A Story, a high-quality dataset of complex writing prompts and human-written stories, and a novel evaluation framework designed specifically for assessing long narratives. We show that Agents' Room generates stories that are preferred by expert evaluators over those produced by baseline systems by leveraging collaboration and specialization to decompose the complex story writing task into tractable components. We provide extensive analysis with automated and human-based metrics of the generated output.

## 2521. ND-SDF: Learning Normal Deflection Fields for High-Fidelity Indoor Reconstruction

链接：https://iclr.cc/virtual/2025/poster/31021 abstract： Neural implicit reconstruction via volume rendering has demonstrated its effectiveness in recovering dense 3D surfaces. However, it is non-trivial to simultaneously recover meticulous geometry and preserve smoothness across regions with differing characteristics. To address this issue, previous methods typically employ geometric priors, which are often constrained by the performance of the prior models. In this paper, we propose ND-SDF, which learns a Normal Deflection field to represent the angular deviation between the scene normal and the prior normal. Unlike previous methods that uniformly apply geometric priors on all samples, introducing significant bias in accuracy, our proposed normal deflection field dynamically learns and adapts the utilization of samples based on their specific characteristics, thereby improving both the accuracy and effectiveness of the model. Our method not only obtains smooth weakly textured regions such as walls and floors but also preserves the geometric details of complex structures. In addition, we introduce a novel ray sampling strategy based on the deflection angle to facilitate the unbiased rendering process, which significantly improves the quality and accuracy of intricate surfaces, especially on thin structures. Consistent improvements on various challenging datasets demonstrate the superiority of our method.

## 2522. LeFusion: Controllable Pathology Synthesis via Lesion-Focused Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/31069 abstract： Patient data from real-world clinical practice often suffers from data scarcity and long-tail imbalances, leading to biased outcomes or algorithmic unfairness. This study addresses these challenges by generating lesion-containing image-segmentation pairs from lesion-free images. Previous efforts in medical imaging synthesis have struggled with separating lesion information from background, resulting in low-quality backgrounds and limited control over the synthetic output. Inspired by diffusion-based image inpainting, we propose LeFusion, a lesion-focused diffusion model. By redesigning the diffusion learning objectives to focus on lesion areas, we simplify the learning process and improve control over the output while preserving high-fidelity backgrounds by integrating forward-diffused background contexts into the reverse diffusion process. Additionally, we tackle two major challenges in lesion texture synthesis: 1) multi-peak and 2) multi-class lesions. We introduce two effective strategies: histogram-based texture control and multi-channel decomposition, enabling the controlled generation of high-quality lesions in difficult scenarios. Furthermore, we incorporate lesion mask diffusion, allowing control over lesion size, location, and boundary, thus increasing lesion diversity. Validated on 3D cardiac lesion MRI and lung nodule CT datasets, LeFusion-generated data significantly improves the performance of state-of-the-art segmentation models, including nnUNet and SwinUNETR.

## 2523. Efficient Active Imitation Learning with Random Network Distillation

链接：https://iclr.cc/virtual/2025/poster/30291 abstract： Developing agents for complex and underspecified tasks, where no clear objective exists, remains challenging but offers many opportunities. This is especially true in video games, where simulated players (bots) need to play realistically, and there is no clear reward to evaluate them. While imitation learning has shown promise in such domains, these methods often fail when agents encounter out-of-distribution scenarios during deployment. Expanding the training dataset is a common solution, but it becomes impractical or costly when relying on human demonstrations. This article addresses active imitation learning, aiming to trigger expert intervention only when necessary, reducing the need for constant expert input along training. We introduce Random Network Distillation DAgger (RND-DAgger), a new active imitation learning method that limits expert querying by using a learned state-based out-of-distribution measure to trigger interventions. This approach avoids frequent expert-agent action comparisons, thus making the expert intervene only when it is useful. We evaluate RND-DAgger against traditional imitation learning and other active approaches in 3D video games (racing and third-person navigation) and in a robotic locomotion task and show that RND-DAgger surpasses previous methods by reducing expert queries.https://sites.google.com/view/rnd-dagger

## 2524. A Computational Framework for Modeling Emergence of Color Vision in the Human Brain

链接：https://iclr.cc/virtual/2025/poster/28836 abstract： It is a mystery how the brain decodes color vision purely from the optic nerve signals it receives, with a core inferential challenge being how it disentangles internal perception with the correct color dimensionality from the unknown encoding properties of the eye. In this paper, we introduce a computational framework for

modeling this emergence of human color vision by simulating both the eye and the cortex. Existing research often overlooks how the cortex develops color vision or represents color space internally, assuming that the color dimensionality is known a priori; however, we argue that the visual cortex has the capability and the challenge of inferring the color dimensionality purely from fluctuations in the optic nerve signals. To validate our theory, we introduce a simulation engine for biological eyes based on established vision science and generate optic nerve signals resulting from looking at natural images. Further, we propose a bio-plausible model of cortical learning based on self-supervised prediction of optic nerve signal fluctuations under natural eye motions. We show that this model naturally learns to generate color vision by disentangling retinal invariants from the sensory signals. When the retina contains $N$ types of color photoreceptors, our simulation shows that $N$-dimensional color vision naturally emerges, verified through formal colorimetry. Using this framework, we also present the first simulation work that successfully boosts the color dimensionality, as observed in gene therapy on squirrel monkeys, and demonstrates the possibility of enhancing human color vision from 3D to 4D.

## 2525. GEVRM: Goal-Expressive Video Generation Model For Robust Visual Manipulation

链接：https://iclr.cc/virtual/2025/poster/28764 abstract： With the rapid development of embodied artificial intelligence, significant progress has been made in vision-language-action (VLA) models for general robot decision-making. However, the majority of existing VLAs fail to account for the inevitable external perturbations encountered during deployment. These perturbations introduce unforeseen state information to the VLA, resulting in inaccurate actions and consequently, a significant decline in generalization performance. The classic internal model control (IMC) principle demonstrates that a closed-loop system with an internal model that includes external input signals can accurately track the reference input and effectively offset the disturbance. We propose a novel closed-loop VLA method GEVRM that integrates the IMC principle to enhance the robustness of robot visual manipulation. The text-guided video generation model in GEVRM can generate highly expressive future visual planning goals. Simultaneously, we evaluate perturbations by simulating responses, which are called internal embeddings and optimized through prototype contrastive learning. This allows the model to implicitly infer and distinguish perturbations from the external environment. The proposed GEVRM achieves state-of-the-art performance on both standard and perturbed CALVIN benchmarks and shows significant improvements in realistic robot tasks.

## 2526. Do Mice Grok? Glimpses of Hidden Progress in Sensory Cortex

链接：https://iclr.cc/virtual/2025/poster/28350 abstract： Does learning of task-relevant representations stop when behavior stops changing? Motivated by recent work in machine learning and the intuitive observation that human experts continue to learn after mastery, we hypothesize that task-specific representation learning in cortex can continue, even when behavior saturates. In a novel reanalysis of recently published neural data, we find evidence for such learning in posterior piriform cortex of mice following continued training on a task, long after behavior saturates at near-ceiling performance ("overtraining"). We demonstrate that class representations in cortex continue to separate during overtraining, so that examples that were incorrectly classified at the beginning of overtraining can abruptly be correctly classified later on, despite no changes in behavior during that time. We hypothesize this hidden learning takes the form of approximate margin maximization; we validate this and other predictions in the neural data, as well as build and interpret a simple synthetic model that recapitulates these phenomena. We conclude by demonstrating how this model of late-time feature learning implies an explanation for the empirical puzzle of overtraining reversal in animal learning, where task-specific representations are more robust to particular task changes because the learned features can be reused.

## 2527. From Sparse Dependence to Sparse Attention: Unveiling How Chain-of-Thought Enhances Transformer Sample Efficiency

链接：https://iclr.cc/virtual/2025/poster/30616 abstract： Chain-of-thought (CoT) significantly enhances the reasoning performance of large language models (LLM). While current theoretical studies often attribute this improvement to increased expressiveness and computational capacity, we argue that expressiveness is not the primary limitation in the LLM regime, as current large models will fail on simple tasks. Using a parity-learning setup, we demonstrate that CoT can substantially improve sample efficiency even when the representation power is sufficient. Specifically, with CoT, a transformer can learn the function within polynomial samples, whereas without CoT, the required sample size is exponential. Additionally, we show that CoT simplifies the learning process by introducing sparse sequential dependencies among input tokens, and leads to a sparse and interpretable attention. We validate our theoretical analysis with both synthetic and real-world experiments, confirming that sparsity in attention layers is a key factor of the improvement induced by CoT.

## 2528. Unsupervised Multiple Kernel Learning for Graphs via Ordinality Preservation

链接：https://iclr.cc/virtual/2025/poster/30865 abstract：

## 2529. Revisiting text-to-image evaluation with Gecko: on metrics, prompts, and human rating

链接：https://iclr.cc/virtual/2025/poster/30151 abstract： While text-to-image (T2I) generative models have become ubiquitous, they do not necessarily generate images that align with a given prompt. While many metrics and benchmarks have been proposed to evaluate T2I models and alignment metrics, the impact of the evaluation components (prompt sets, human annotations, evaluation task) has not been systematically measured.We find that looking at only *one slice of data*, i.e. one set of capabilities or human annotations, is not enough to obtain stable conclusions that generalise to new conditions or slices when evaluating T2I models or alignment metrics. We address this by introducing an evaluation suite of $>$100K annotations across four human annotation templates that comprehensively evaluates models' capabilities across a range of methods for gathering human annotations and comparing models.In particular, we propose (1) a carefully curated set of prompts -- *Gecko2K*; (2) a statistically grounded method of comparing T2I models; and (3) how to systematically evaluate metrics under three *evaluation tasks -- model ordering, pair-wise instance scoring, point-wise instance scoring*.Using this evaluation suite, we evaluate a wide range of metrics and find that a metric may do better in one setting but worse in another.As a result, we introduce a new, interpretable auto-eval metric that is consistently better correlated with human ratings than such existing metrics on our evaluation suite--across different human templates and evaluation settings--and on TIFA160.

## 2530. Mitigating Information Loss in Tree-Based Reinforcement Learning via Direct Optimization

链接：https://iclr.cc/virtual/2025/poster/28234 abstract： Reinforcement learning (RL) has seen significant success across various domains, but its adoption is often limited by the black-box nature of neural network policies, making them difficult to interpret. In contrast, symbolic policies allow representing decision-making strategies in a compact and interpretable way. However, learning symbolic policies directly within on-policy methods remains challenging.In this paper, we introduce SYMPOL, a novel method for SYMbolic tree-based on-POLicy RL. SYMPOL employs a tree-based model integrated with a policy gradient method, enabling the agent to learn and adapt its actions while maintaining a high level of interpretability.We evaluate SYMPOL on a set of benchmark RL tasks, demonstrating its superiority over alternative tree-based RL approaches in terms of performance and interpretability. Unlike existing methods, it enables gradient-based, end-to-end learning of interpretable, axis-aligned decision trees within standard on-policy RL algorithms. Therefore, SYMPOL can become the foundation for a new class of interpretable RL based on decision trees. Our implementation is available under: https://github.com/s-marton/sympol

## 2531. Salvage: Shapley-distribution Approximation Learning Via Attribution Guided Exploration for Explainable Image Classification

链接：https://iclr.cc/virtual/2025/poster/29378 abstract： The integration of deep learning into critical vision application areas has given rise to a necessity for techniques that can explain the rationale behind predictions. In this paper, we address this need by introducing Salvage, a novel removal-based explainability method for image classification. Our approach involves training an explainer model that learns the prediction distribution of the classifier on masked images. We first introduce the concept of Shapley-distributions, which offers a more accurate approximation of classification probability distributions than existing methods. Furthermore, we address the issue of unbalanced important and unimportant features. In such settings, naive uniform sampling of feature subsets often results in a highly unbalanced ratio of samples with high and low prediction likelihoods, which can hinder effective learning. To mitigate this, we propose an informed sampling strategy that leverages approximated feature importance scores, thereby reducing imbalance and facilitating the estimation of underrepresented features. After incorporating these two principles into our method, we conducted an extensive analysis on the ImageNette, MURA, WBC, and Pet datasets. The results show that Salvage outperforms various baseline explainability methods, including attention-, gradient-, and removal-based approaches, both qualitatively and quantitatively. Furthermore, we demonstrate that our explainer model can serve as a fully explainable classifier without a major decrease in classification performance, paving the way for fully explainable image classification.

## 2532. Collaborative Discrete-Continuous Black-Box Prompt Learning for Language Models

链接：https://iclr.cc/virtual/2025/poster/28111 abstract： Large Scale Pre-Trained Language Models (PTMs) have demonstrated unprecedented capabilities across diverse natural language processing tasks. Adapting such models to downstream tasks is computationally intensive and time-consuming, particularly in black-box scenarios common in Language-Model-as-a-Service (LMaaS) environments, where model parameters and gradients are inaccessible. Recently, black-box prompt learning using zeroth-order gradients has emerged as a promising approach to address these challenges by optimizing learnable continuous prompts in embedding spaces, starting with \textit{randomly initialized discrete text prompts}. However, its reliance on randomly initialized discrete prompts limits adaptability to diverse downstream tasks or models. To address this limitation,this paper introduces ZO-PoG, a novel framework that optimizes prompts through a collaborative approach, combining Policy Gradient optimization for initial discrete text prompts and Zeroth-Order optimization for continuous prompts in embedding space. By optimizing collaboratively between discrete and continuous prompts, ZO-PoG maximizes adaptability to downstream tasks, achieving superior results without direct access to the model’s internal structures.Importantly, we establish the sub-linear convergence of ZO-PoG under mild assumptions.The experiments on different datasets demonstrate significant improvements in various tasks compared to the baselines.

## 2533. Overcoming Lower-Level Constraints in Bilevel Optimization: A Novel

## Approach with Regularized Gap Functions

链接：https://iclr.cc/virtual/2025/poster/29020 abstract： Constrained bilevel optimization tackles nested structures present in constrained learning tasks like constrained meta-learning, adversarial learning, and distributed bilevel optimization. However, existing bilevel optimization methods mostly are typically restricted to specific constraint settings, such as linear lower-level constraints. In this work, we overcome this limitation and develop a new single-loop, Hessian-free constrained bilevel algorithm capable of handling more general lower-level constraints. We achieve this by employing a doubly regularized gap function tailored to the constrained lower-level problem, transforming constrained bilevel optimization into an equivalent single-level optimization problem with a single smooth constraint. We rigorously establish the non-asymptotic convergence analysis of the proposed algorithm under the convexity of lower-level problem, avoiding the need for strong convexity assumptions on the lower-level objective or coupling convexity assumptions on lower-level constraints found in existing literature. Additionally, the generality of our method allows for its extension to bilevel optimization with minimax lower-level problem. We evaluate the effectiveness and efficiency of our algorithm on various synthetic problems, typical hyperparameter learning tasks, and generative adversarial network.

## 2534. Adversarial Training for Defense Against Label Poisoning Attacks

链接：https://iclr.cc/virtual/2025/poster/29458 abstract： As machine learning models grow in complexity and increasingly rely on publicly sourced data, such as the human-annotated labels used in training large language models, they become more vulnerable to label poisoning attacks.These attacks, in which adversaries subtly alter the labels within a training dataset, can severely degrade model performance, posing significant risks in critical applications.In this paper, we propose $\textbf{Floral}$, a novel adversarial training defense strategy based on support vector machines (SVMs) to counter these threats. Utilizing a bilevel optimization framework, we cast the training process as a non-zero-sum Stackelberg game between an $\textit{attacker}$, who strategically poisons critical training labels, and the $\textit{model}$, which seeks to recover from such attacks. Our approach accommodates various model architectures and employs a projected gradient descent algorithm with kernel SVMs for adversarial training. We provide a theoretical analysis of our algorithm’s convergence properties and empirically evaluate $\textbf{Floral}$'s effectiveness across diverse classification tasks.Compared to robust baselines and foundation models such as RoBERTa, $\textbf{Floral}$ consistently achieves higher robust accuracy under increasing attacker budgets.These results underscore the potential of $\textbf{Floral}$ to enhance the resilience of machine learning models against label poisoning threats, thereby ensuring robust classification in adversarial settings.

## 2535. qNBO: quasi-Newton Meets Bilevel Optimization

链接：https://iclr.cc/virtual/2025/poster/30570 abstract： Bilevel optimization, which addresses challenges in hierarchical learning tasks, has gained significant interest in machine learning. Implementing gradient descent for bilevel optimization presents computational hurdles, notably the need to compute the exact lower-level solution and the inverse Hessian of the lower-level objective. While these two aspects are inherently connected, existing methods typically handle them separately by solving the lower-level problem and a linear system for the inverse Hessian-vector product. In this paper, we introduce a general framework to tackle these computational challenges in a coordinated manner. Specifically, we leverage quasi-Newton algorithms to accelerate the solution of the lower-level problem while efficiently approximating the inverse Hessian-vector product. Furthermore, by leveraging the superlinear convergence properties of BFGS, we establish a non-asymptotic convergence analysis for the BFGS adaptation within our framework. Numerical experiments demonstrate the comparable or superior performance of our proposed algorithms in real-world learning tasks, including hyperparameter optimization, data hyper-cleaning, and few-shot meta-learning.

## 2536. Faster Inference of Flow-Based Generative Models via Improved Data-Noise Coupling

链接：https://iclr.cc/virtual/2025/poster/28167 abstract： Conditional Flow Matching (CFM), a simulation-free method for training continuous normalizing flows, provides an efficient alternative to diffusion models for key tasks like image and video generation. The performance of CFM in solving these tasks depends on the way data is coupled with noise. A recent approach uses minibatch optimal transport (OT) to reassign noise-data pairs in each training step to streamline sampling trajectories and thus accelerate inference. However, its optimization is restricted to individual minibatches, limiting its effectiveness on large datasets. To address this shortcoming, we introduce LOOM-CFM (Looking Out Of Minibatch-CFM), a novel method to extend the scope of minibatch OT by preserving and optimizing these assignments across minibatches over training time. Our approach demonstrates consistent improvements in the sampling speed-quality trade-off across multiple datasets. LOOM-CFM also enhances distillation initialization and supports high-resolution synthesis in latent space training.

## 2537. Generalizable Human Gaussians from Single-View Image

链接：https://iclr.cc/virtual/2025/poster/28987 abstract： In this work, we tackle the task of learning 3D human Gaussians from a single image, focusing on recovering detailed appearance and geometry including unobserved regions. We introduce a single-view generalizable Human Gaussian Model (HGM), which employs a novel generate-then-refine pipeline with the guidance from human body prior and diffusion prior. Our approach uses a ControlNet to refine rendered back-view images from coarse predicted human Gaussians, then uses the refined image along with the input image to reconstruct refined human Gaussians. To

mitigate the potential generation of unrealistic human poses and shapes, we incorporate human priors from the SMPL-X model as a dual branch, propagating image features from the SMPL-X volume to the image Gaussians using sparse convolution and attention mechanisms. Given that the initial SMPL-X estimation might be inaccurate, we gradually refine it with our HGM model. We validate our approach on several publicly available datasets. Our method surpasses previous methods in both novel view synthesis and surface reconstruction. Our approach also exhibits strong generalization for cross-dataset evaluation and in-the-wild images.

## 2538. Leveraging Flatness to Improve Information-Theoretic Generalization Bounds for SGD

链接：https://iclr.cc/virtual/2025/poster/28297 abstract： Information-theoretic (IT) generalization bounds have been used to study the generalization of learning algorithms. These bounds are intrinsically data- and algorithm-dependent so that one can exploit the properties of data and algorithm to derive tighter bounds. However, we observe that although the flatness bias is crucial for SGD's generalization, these bounds fail to capture the improved generalization under better flatness and are also numerically loose. This is caused by the inadequate leverage of SGD's flatness bias in existing IT bounds. This paper derives a more flatness-leveraging IT bound for the flatness-favoring SGD. The bound indicates the learned models generalize better if the large-variance directions of the final weight covariance have small local curvatures in the loss landscape. Experiments on deep neural networks show our bound not only correctly reflects the better generalization when flatness is improved, but is also numerically much tighter. This is achieved by a flexible technique called "omniscient trajectory". When applied to Gradient Descent's minimax excess risk on convex-Lipschitz-Bounded problems, it improves representative IT bounds' $\Omega(1)$ rates to $O(1/\sqrt{n})$. It also implies a by-pass of memorization-generalization trade-offs. Codes are available at https://github.com/peng-ze/omniscient-bounds.

## 2539. Certified Robustness Under Bounded Levenshtein Distance

链接：https://iclr.cc/virtual/2025/poster/29038 abstract：

## 2540. Block-Attention for Efficient Prefilling

链接：https://iclr.cc/virtual/2025/poster/30791 abstract： We introduce Block-attention, an attention mechanism designed to address the increased inference latency and cost in Retrieval-Augmented Generation (RAG) scenarios. Traditional approaches often encode the entire context in an auto-regressive manner.Instead, Block-attention divides retrieved documents into discrete blocks, with each block independently calculating key-value (KV) states except for the final block.In RAG scenarios, by defining each passage as a block, Block-attention enables us to reuse the KV states of passages that have been seen before, thereby significantly reducing the latency and the computation overhead during inference.The implementation of Block-attention involves block segmentation, position re-encoding, and fine-tuning the LLM to adapt to the Block-attention mechanism. Experiments on 11 diverse benchmarks, including RAG, ICL, and general domains, demonstrate that after block fine-tuning, the Block-attention model not only achieves performance comparable to that of full-attention models, but can also seamlessly switch between the block and full attention modes without any performance loss.Notably, Block-attention significantly reduces the time to first token (TTFT) and floating point operations (FLOPs) to a very low level. It only takes 45 ms to output the first token for an input sequence with a total length of 32K. Compared to the full-attention models, the TTFT and corresponding FLOPs are reduced by 98.7\% and 99.8\%, respectively. Additionally, in Appendix A, we elaborate on how Block-attention is applied in Game AI scenario and the substantial potential benefits it entails. We strongly suggest researchers in the gaming field not to overlook this section.

## 2541. Efficient Interpolation between Extragradient and Proximal Methods for Weak MVIs

链接：https://iclr.cc/virtual/2025/poster/29271 abstract： We study nonmonotone games satisfying the weak Minty variational inequality (MVI) with parameter $\rho \in (-\tfrac{1}{L}, \infty)$, where $L$ is the Lipschitz constant of the gradient operator. An error corrected version of the inexact proximal point algorithm is proposed, with which we establish the first $\mathcal O(1/\epsilon)$ rate for the entire range $\rho \in (-\tfrac{1}{L}, \infty)$, thus removing a logarithmic factor compared with the complexity of existing methods. The scheme automatically selects the needed accuracy for the proximal computation, and can recover the relaxed extragradient method when $\rho > -\tfrac{1}{2L}$ and the relaxed proximal point algorithm (rPPA) when $\rho > -\tfrac{1}{L}$. Due to the error correction, the scheme inherits the strong properties of the *exact* rPPA. Specifically, we show that linear convergence is automatically achieved under appropriate conditions. Tightness for the range of $\rho$ is established through a lower bound for rPPA. Central to the algorithmic construction is a halfspace projection, where the key insight is that the allowed error tolerance can both be used to correct for the proximal approximation and to enlarge the problem class.

## 2542. Rotated Runtime Smooth: Training-Free Activation Smoother for accurate INT4 inference

链接：https://iclr.cc/virtual/2025/poster/29374 abstract： Large language models have demonstrated promising capabilities

upon scaling up parameters. However, serving large language models incurs substantial computation and memory movement costs due to their large scale. Quantization methods have been employed to reduce service costs and latency. Nevertheless, outliers in activations hinder the development of INT4 weight-activation quantization. Existing approaches separate outliers and normal values into two matrices or migrate outliers from activations to weights, suffering from high latency or accuracy degradation. Based on observing activations from large language models, outliers can be classified into channel-wise and spike outliers.In this work, we propose Rotated Runtime Smooth (RRS), a plug-and-play activation smoother for quantization, consisting of Runtime Smooth and the Rotation operation. Runtime Smooth (RS) is introduced to eliminate channel-wise outliers by smoothing activations with channel-wise maximums during runtime. The Rotation operation can narrow the gap between spike outliers and normal values, alleviating the effect of victims caused by channel-wise smoothing.The proposed method outperforms the state-of-the-art method in the LLaMA and Qwen families and improves WikiText-2 perplexity from 57.33 to 6.66 for INT4 inference.

## 2543. TODO: Enhancing LLM Alignment with Ternary Preferences

链接：https://iclr.cc/virtual/2025/poster/27951 abstract： Aligning large language models (LLMs) with human intent is critical for enhancing their performance across a variety of tasks. Standard alignment techniques, such as Direct Preference Optimization (DPO), often rely on the binary Bradley-Terry (BT) model, which can struggle to capture the complexities of human preferences—particularly in the presence of noisy or inconsistent labels and frequent ties. To address these limitations, we introduce the Tie-rank Oriented Bradley-Terry model (TOBT), an extension of the BT model that explicitly incorporates ties, enabling more nuanced preference representation. Building on this, we propose Tie-rank Oriented Direct Preference Optimization (TODO), a novel alignment algorithm that leverages TOBT's ternary ranking system to improve preference alignment. In evaluations on Mistral-7B and Llama 3-8B models, TODO consistently outperforms DPO in modeling preferences across both in-distribution and out-of-distribution datasets. Additional assessments using MT Bench and benchmarks such as Piqa, ARC-c, and MMLU further demonstrate TODO's superior alignment performance. Notably, TODO also shows strong results in binary preference alignment, highlighting its versatility and potential for broader integration into LLM alignment. The code for TODO is made publicly available.

## 2544. Learning Dynamics of LLM Finetuning

链接：https://iclr.cc/virtual/2025/poster/28056 abstract： Learning dynamics, which describes how the learning of specific training examples influences the model's predictions on other examples, gives us a powerful tool for understanding the behavior of deep learning systems. We study the learning dynamics of large language models during different types of finetuning, by analyzing the step-wise decomposition of how influence accumulates among different potential responses. Our framework allows a uniform interpretation of many interesting observations about the training of popular algorithms for both instruction tuning and preference tuning. In particular, we propose a hypothetical explanation of why specific types of hallucination are strengthened after finetuning, e.g., the model might use phrases or facts in the response for question B to answer question A, or the model might keep repeating similar simple phrases when generating responses. We also extend our framework and highlight a unique ``squeezing effect'' to explain a previously observed phenomenon in off-policy direct preference optimization (DPO), where running DPO for too long makes even the desired outputs less likely. This framework also provides insights into where the benefits of on-policy DPO and other variants come from. The analysis not only provides a novel perspective of understanding LLM's finetuning but also inspires a simple, effective method to improve alignment performance.

## 2545. Two Effects, One Trigger: On the Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/27997 abstract： Contrastive vision-language models (VLMs), like CLIP, have gained popularity for their versatile applicability to various downstream tasks. Despite their successes in some tasks, like zero-shot object recognition, they perform surprisingly poor on other tasks, like attribute recognition. Previous work has attributed these challenges to the modality gap, a separation of image and text in the shared representation space, and to a bias towards objects over other factors, such as attributes. In this analysis paper, we investigate both phenomena thoroughly. We evaluated off-the-shelf VLMs and while the gap's influence on performance is typically overshadowed by other factors, we find indications that closing the gap indeed leads to improvements. Moreover, we find that, contrary to intuition, only few embedding dimensions drive the gap and that the embedding spaces are differently organized. To allow for a clean study of object bias, we introduce a definition and a corresponding measure of it. Equipped with this tool, we find that object bias does not lead to worse performance on other concepts, such as attributes per se. However, why do both phenomena, modality gap and object bias, emerge in the first place? To answer this fundamental question and uncover some of the inner workings of contrastive VLMs, we conducted experiments that allowed us to control the amount of shared information between the modalities. These experiments revealed that the driving factor behind both the modality gap and the object bias, is an information imbalance between images and captions, and unveiled an intriguing connection between the modality gap and entropy of the logits.

## 2546. Injective flows for star-like manifolds

链接：https://iclr.cc/virtual/2025/poster/30082 abstract： Normalizing Flows (NFs) are powerful and efficient models for density estimation. When modeling densities on manifolds, NFs can be generalized to injective flows but the Jacobian determinant becomes computationally prohibitive. Current approaches either consider bounds on the log-likelihood or rely on some approximations of the Jacobian determinant. In contrast, we propose injective flows for star-like manifolds and show that for such

manifolds we can compute the Jacobian determinant exactly and efficiently. This aspect is particularly relevant for variational inference settings, where no samples are available and only some unnormalized target is known. Among many, we showcase the relevance of modeling densities on star-like manifolds in two settings. Firstly, we introduce a novel Objective Bayesian approach for penalized likelihood models by interpreting level-sets of the penalty as star-like manifolds. Secondly, we consider probabilistic mixing models and introduce a general method for variational inference by defining the posterior of mixture weights on the probability simplex.

## 2547. Exploring Prosocial Irrationality for LLM Agents: A Social Cognition View

链接：https://iclr.cc/virtual/2025/poster/27998 abstract： Large language models (LLMs) have been shown to face hallucination issues due to the data they trained on often containing human bias; whether this is reflected in the decision-making process of LLM agents remains under-explored. As LLM Agents are increasingly employed in intricate social environments, a pressing and natural question emerges: Can we utilize LLM Agents' systematic hallucinations to mirror human cognitive biases, thus exhibiting irrational social intelligence? In this paper, we probe the irrational behavior among contemporary LLM agents by melding practical social science experiments with theoretical insights. Specifically, we propose CogMir, an open-ended Multi-LLM Agents framework that utilizes hallucination properties to assess and enhance LLM Agents' social intelligence through cognitive biases. Experimental results on CogMir subsets show that LLM Agents and humans exhibit high consistency in irrational and prosocial decision-making under uncertain conditions, underscoring the prosociality of LLM Agents as social entities and highlighting the significance of hallucination properties. Additionally, CogMir framework demonstrates its potential as a valuable platform for encouraging more research into the social intelligence of LLM Agents.

## 2548. Causally Motivated Sycophancy Mitigation for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/27727 abstract： Incorporating user preferences into large language models (LLMs) can enhance the personalization and reliability of model outputs and facilitate the application of LLMs to real-world scenarios. However, leveraging user preferences can be a double-edged sword. Recent studies have found that improper utilization can incur sycophancy, where LLMs prioritize alignment with user preferences over the correctness of their outputs. To address sycophancy in LLMs, we analyze and model the problem through the lens of structured causal models (SCMs). We attribute sycophancy to LLMs' reliance on spurious correlations between user preferences and model outputs in this paper. Based on the proposed SCMs, we develop a novel framework, termed CAUSM, to mitigate sycophancy in LLMs by exploiting a significant causal signature. Specifically, we eliminate the spurious correlations embedded in the intermediate layers of LLMs through causally motivated head reweighting, and then calibrate the intra-head knowledge along the causal representation direction. Extensive experiments are conducted across diverse language tasks to demonstrate the superiority of our method over state-of-the-art competitors in mitigating sycophancy in LLMs.

## 2549. VideoShield: Regulating Diffusion-based Video Generation Models via Watermarking

链接：https://iclr.cc/virtual/2025/poster/27941 abstract：

## 2550. Flaws of ImageNet, Computer Vision's Favourite Dataset

链接：https://iclr.cc/virtual/2025/poster/31342 abstract： Since its release, ImageNet-1k dataset has become a gold standard for evaluating model performance. It has served as the foundation for numerous other datasets and training tasks in computer vision. As models have improved in accuracy, issues related to label correctness have become increasingly apparent. In this blog post, we analyze the issues in the ImageNet-1k dataset, including incorrect labels, overlapping or ambiguous class definitions, training-evaluation domain shifts, and image duplicates. The solutions for some problems are straightforward. For others, we hope to start a broader conversation about refining this influential dataset to better serve future research.

## 2551. Influence Functions for Scalable Data Attribution in Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/28904 abstract： Diffusion models have led to significant advancements in generative modelling. Yet their widespread adoption poses challenges regarding data attribution and interpretability. In this paper, we aim to help address such challenges in diffusion models by extending influence functions. Influence function-based data attribution methods approximate how a model's output would have changed if some training data were removed. In supervised learning, this is usually used for predicting how the loss on a particular example would change. For diffusion models, we focus on predicting the change in the probability of generating a particular example via several proxy measurements. We show how to formulate influence functions for such quantities and how previously proposed methods can be interpreted as particular design choices in our framework. To ensure scalability of the Hessian computations in influence functions, we use a K-FAC approximation based on generalised Gauss-Newton matrices specifically tailored to diffusion models. We show that our recommended method outperforms previously proposed data attribution methods on common data attribution evaluations, such as the Linear Data-modelling Score (LDS) or retraining without top influences, without the need for method-specific hyperparameter tuning.

## 2552. HShare: Fast LLM Decoding by Hierarchical Key-Value Sharing

链接：https://iclr.cc/virtual/2025/poster/29532 abstract： The frequent retrieval of Key-Value (KV) cache data has emerged as a significant factor contributing to the inefficiency of the inference process in large language models. Previous research has demonstrated that a small subset of critical KV cache tokens largely influences attention outcomes, leading to methods that either employ fixed sparsity patterns or dynamically select critical tokens based on the query. While dynamic sparse patterns have proven to be more effective, they introduce significant computational overhead, as critical tokens must be reselected for each self-attention computation. In this paper, we reveal substantial similarities in KV cache token criticality across neighboring queries, layers, and heads. Motivated by this insight, we propose HShare, a hierarchical KV sharing framework. HShare facilitates the sharing of critical KV cache token indices across layers, heads, and queries, which significantly reduces the computational overhead associated with query-aware dynamic token sparsity. In addition, we introduce a greedy algorithm that dynamically determines the optimal layer-level and head-level sharing configuration for the decoding phase. We evaluate the effectiveness and efficiency of HShare across various tasks using three models: LLaMA2-7b, LLaMA3-70b, and Mistral-7b. Experimental results demonstrate that HShare achieves competitive accuracy with different sharing ratios, while delivering up to an $8.6\times$ speedup in self-attention operations and a $2.7\times$ improvement in end-to-end throughput compared with FlashAttention2 and GPT-fast respectively. The source code is publicly available at ~\url{https://github.com/wuhuaijin/HShare}.

## 2553. ACE: All-round Creator and Editor Following Instructions via Diffusion Transformer

链接：https://iclr.cc/virtual/2025/poster/30543 abstract： Diffusion models have emerged as a powerful generative technology and have been found to be applicable in various scenarios. Most existing foundational diffusion models are primarily designed for text-guided visual generation and do not support multi-modal conditions, which are essential for many visual editing tasks. This limitation prevents these foundational diffusion models from serving as a unified model in the field of visual generation, like GPT-4 in the natural language processing field. In this work, we propose ACE, an All-round Creator and Editor, which achieves comparable performance compared to those expert models in a wide range of visual generation tasks. To achieve this goal, we first introduce a unified condition format termed Long-context Condition Unit (LCU), and propose a novel Transformer-based diffusion model that uses LCU as input, aiming for joint training across various generation and editing tasks. Furthermore, we propose an efficient data collection approach to address the issue of the absence of available training data. It involves acquiring pairwise images with synthesis-based or clustering-based pipelines and supplying these pairs with accurate textual instructions by leveraging a fine-tuned multi-modal large language model. To comprehensively evaluate the performance of our model, we establish a benchmark of manually annotated pairs data across a variety of visual generation tasks. The extensive experimental results demonstrate the superiority of our model in visual generation fields. Thanks to the all-in-one capabilities of our model, we can easily build a multi-modal chat system that responds to any interactive request for image creation using a single model to serve as the backend, avoiding the cumbersome pipeline typically employed in visual agents.

## 2554. Indirect Gradient Matching for Adversarial Robust Distillation

链接：https://iclr.cc/virtual/2025/poster/28611 abstract： Adversarial training significantly improves adversarial robustness, but superior performance is primarily attained with large models. This substantial performance gap for smaller models has spurred active research into adversarial distillation (AD) to mitigate the difference. Existing AD methods leverage the teacher’s logits as a guide.In contrast to these approaches, we aim to transfer another piece of knowledge from the teacher, the input gradient.In this paper, we propose a distillation module termed Indirect Gradient Distillation Module (IGDM) that indirectly matches the student’s input gradient with that of the teacher.Experimental results show that IGDM seamlessly integrates with existing AD methods, significantly enhancing their performance.Particularly, utilizing IGDM on the CIFAR-100 dataset improves the AutoAttack accuracy from 28.06\% to 30.32\% with the ResNet-18 architecture and from 26.18\% to 29.32\% with the MobileNetV2 architecture when integrated into the SOTA method without additional data augmentation.

## 2555. Streamlining Redundant Layers to Compress Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30185 abstract： This paper introduces LLM-Streamline, a pioneer work on layer pruning for large language models (LLMs). It is based on the observation that different layers have varying impacts on hidden states, enabling the identification of less important layers to be pruned. LLM-Streamline comprises two parts: layer pruning, which removes consecutive layers with the lowest importance based on target sparsity, and layer replacement, a novel module that trains a lightweight network to replace the pruned layers to mitigate performance loss. Additionally, a new metric called stability is proposed to address the limitations of the widely used accuracy metric in evaluating model compression. Experiments show that LLM-Streamline outperforms both previous and concurrent state-of-the-art pruning methods in terms of both performance and training efficiency. Our code is available at \href{https://github.com/RUCKBReasoning/LLM-Streamline} {this repository}.

## 2556. Lossy Compression with Pretrained Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/28188 abstract： We apply Theis et al. (2022)'s DiffC algorithm to Stable Diffusion 1.5, 2.1, XL, and and Flux-dev, and demonstrate that these pretrained models are remarkably capable lossy image compressors. A principled algorithm for compression using pretrained diffusion models has been understood since at least 2020 (Ho et al.), but

challenges in reverse-channel coding have prevented such algorithms from ever being fully implemented. We introduce simple workarounds that lead to the first complete implementation of DiffC, which is capable of compressing and decompressing images using Stable Diffusion in under 10 seconds. Despite requiring no additional training, our method is competitive with other state-of-the-art generative compression methods at low ultra-low bitrates.

## 2557. DiscoveryBench: Towards Data-Driven Discovery with Large Language Models

链接：https://iclr.cc/virtual/2025/poster/32050 abstract： Can the rapid advances in code generation, function calling, and data analysis using large language models (LLMs) help automate the search and verification of hypotheses purely from a set of provided datasets? To evaluate this question, we present DiscoveryBench, the first comprehensive benchmark that formalizes the multi-step process of data-driven discovery. The benchmark is designed to systematically assess current model capabilities in discovery tasks and provide a useful resource for improving them. Our benchmark contains 264 tasks collected across 6 diverse domains, such as sociology and engineering, by manually deriving discovery workflows from published papers to approximate the real-world challenges faced by researchers, where each task is defined by a dataset, its metadata, and a discovery goal in natural language. We additionally provide 903 synthetic tasks to conduct controlled evaluations on data-driven workflows that are not covered in the manually collected split. Furthermore, our structured formalism of data-driven discovery enables a facet-based evaluation that provides useful insights into different failure modes. We evaluate several popular LLM-based reasoning frameworks using both open and closed LLMs as baselines on DiscoveryBench and find that even the best system scores only 25%. Our benchmark, thus, illustrates the challenges in autonomous data-driven discovery and serves as a valuable resource for the community to make progress.

## 2558. Finally Rank-Breaking Conquers MNL Bandits: Optimal and Efficient Algorithms for MNL Assortment

链接：https://iclr.cc/virtual/2025/poster/28552 abstract： We address the problem of active online assortment optimization problem with preference feedback, which is a framework for modeling user choices and subsetwise utility maximization. The framework is useful in various real-world applications including ad placement, online retail, recommender systems, and fine-tuning language models, amongst many others. The problem, although has been studied in the past, lacks an intuitive and practical solution approach with simultaneously efficient algorithm and optimal regret guarantee. E.g., popularly used assortment selection algorithms often require the presence of a `strong reference" which is always included in the choice sets, further they are also designed to offer the same assortments repeatedly until the reference item gets selected---all such requirements are quite unrealistic for practical applications. In this paper, we designed efficient algorithms for the problem of regret minimization in assortment selection with \emph{Plackett Luce} (PL) based user choices. We designed a novel concentration guarantee for estimating the score parameters of the PL model using\emph{Pairwise Rank-Breaking}', which builds the foundation of our proposed algorithms. Moreover, our methods are practical, provably optimal, and devoid of the aforementioned limitations of the existing methods.

## 2559. Multi-agent cooperation through learning-aware policy gradients

链接：https://iclr.cc/virtual/2025/poster/30261 abstract： Self-interested individuals often fail to cooperate, posing a fundamental challenge for multi-agent learning. How can we achieve cooperation among self-interested, independent learning agents? Promising recent work has shown that in certain tasks cooperation can be established between ``learning-aware" agents who model the learning dynamics of each other. Here, we present the first unbiased, higher-derivative-free policy gradient algorithm for learning-aware reinforcement learning, which takes into account that other agents are themselves learning through trial and error based on multiple noisy trials. We then leverage efficient sequence models to condition behavior on long observation histories that contain traces of the learning dynamics of other agents. Training long-context policies with our algorithm leads to cooperative behavior and high returns on standard social dilemmas, including a challenging environment where temporally-extended action coordination is required. Finally, we derive from the iterated prisoner's dilemma a novel explanation for how and when cooperation arises among self-interested learning-aware agents.

## 2560. Highly Efficient Self-Adaptive Reward Shaping for Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29692 abstract： Reward shaping is a reinforcement learning technique that addresses the sparse-reward problem by providing frequent, informative feedback. We propose an efficient self-adaptive reward-shaping mechanism that uses success rates derived from historical experiences as shaped rewards. The success rates are sampled from Beta distributions, which evolve from uncertainty to reliability as data accumulates. Initially, shaped rewards are stochastic to encourage exploration, gradually becoming more certain to promote exploitation and maintain a natural balance between exploration and exploitation. We apply Kernel Density Estimation (KDE) with Random Fourier Features (RFF) to derive Beta distributions, providing a computationally efficient solution for continuous and high-dimensional state spaces. Our method, validated on tasks with extremely sparse rewards, improves sample efficiency and convergence stability over relevant baselines.

## 2561. SageAttention: Accurate 8-Bit Attention for Plug-and-play Inference

## Acceleration

链接：https://iclr.cc/virtual/2025/poster/29829 abstract： The transformer architecture predominates across various models. As the heart of the transformer, attention has a computational complexity of $O(N^2)$, compared to $O(N)$ for linear transformations. When handling large sequence lengths, attention becomes the primary time-consuming component. Although quantization has proven to be an effective method for accelerating model inference, existing quantization methods primarily focus on optimizing the linear layer.In response, we first analyze the feasibility of quantization in attention detailedly. Following that, we propose SageAttention, a highly efficient and accurate quantization method for attention. The OPS (operations per second) of our approach outperforms FlashAttention2 and xformers by about 2.1x and 2.7x, respectively. SageAttention also achieves superior accuracy performance over FlashAttention3. Comprehensive experiments confirm that our approach incurs almost no end-to-end metrics loss across diverse models—including those for large language processing, image generation, and video generation. The code is available at https://github.com/thu-ml/SageAttention.

## 2562. GeSubNet: Gene Interaction Inference for Disease Subtype Network Generation

链接：https://iclr.cc/virtual/2025/poster/28630 abstract： Retrieving gene functional networks from knowledge databases presents a challenge due to the mismatch between disease networks and subtype-specific variations. Current solutions, including statistical and deep learning methods, often fail to effectively integrate gene interaction knowledge from databases or explicitly learn subtype-specific interactions. To address this mismatch, we propose GeSubNet, which learns a unified representation capable of predicting gene interactions while distinguishing between different disease subtypes. Graphs generated by such representations can be considered subtype-specific networks. GeSubNet is a multi-step representation learning framework with three modules: First, a deep generative model learns distinct disease subtypes from patient gene expression profiles. Second, a graph neural network captures representations of prior gene networks from knowledge databases, ensuring accurate physical gene interactions. Finally, we integrate these two representations using an inference loss that leverages graph generation capabilities, conditioned on the patient separation loss, to refine subtype-specific information in the learned representation. GeSubNet consistently outperforms traditional methods, with average improvements of 30.6%, 21.0%, 20.1%, and 56.6% across four graph evaluation metrics, averaged over four cancer datasets. Particularly, we conduct a biological simulation experiment to assess how the behavior of selected genes from over 11,000 candidates affects subtypes or patient distributions. The results show that the generated network has the potential to identify subtype-specific genes with an 83% likelihood of impacting patient distribution shifts.

## 2563. The Pitfalls of Memorization: When Memorization Hurts Generalization

链接：https://iclr.cc/virtual/2025/poster/27908 abstract： Neural networks often learn simple explanations that fit the majority of the data while memorizing exceptions that deviate from these explanations. This behavior leads to poor generalization when the learned explanations rely on spurious correlations. In this work, we formalize $\textit{the interplay between memorization and generalization}$, showing that spurious correlations would particularly lead to poor generalization when are combined with memorization. Memorization can reduce training loss to zero, leaving no incentive to learn robust, generalizable patterns. To address this, we propose $\textit{memorization-aware training}$ (MAT), which uses held-out predictions as a signal of memorization to shift a model's logits. MAT encourages learning robust patterns invariant across distributions, improving generalization under distribution shifts.

## 2564. IPDreamer: Appearance-Controllable 3D Object Generation with Complex Image Prompts

链接：https://iclr.cc/virtual/2025/poster/31077 abstract： Recent advances in 3D generation have been remarkable, with methods such as DreamFusion leveraging large-scale text-to-image diffusion-based models to guide 3D object generation. These methods enable the synthesis of detailed and photorealistic textured objects. However, the appearance of 3D objects produced by such text-to-3D models is often unpredictable, and it is hard for single-image-to-3D methods to deal with images lacking a clear subject, complicating the generation of appearance-controllable 3D objects from complex images. To address these challenges, we present IPDreamer, a novel method that captures intricate appearance features from complex Image Prompts and aligns the synthesized 3D object with these extracted features, enabling high-fidelity, appearance-controllable 3D object generation. Our experiments demonstrate that IPDreamer consistently generates high-quality 3D objects that align with both the textual and complex image prompts, highlighting its promising capability in appearance-controlled, complex 3D object generation.

## 2565. GeoX: Geometric Problem Solving Through Unified Formalized Vision-Language Pre-training

链接：https://iclr.cc/virtual/2025/poster/30888 abstract： Despite their proficiency in general tasks, Multi-modal Large Language Models (MLLMs) struggle with automatic Geometry Problem Solving (GPS), which demands understanding diagrams, interpreting symbols, and performing complex reasoning. This limitation arises from their pre-training on natural images and texts, along with the lack of automated verification in the problem-solving process. Besides, current geometric

specialists are limited by their task-specific designs, making them less effective for broader geometric problems. To this end, we present GeoX, a multi-modal large model focusing on geometric understanding and reasoning tasks. Given the significant differences between geometric diagram-symbol and natural image-text, we introduce unimodal pre-training to develop a diagram encoder and symbol decoder, enhancing the understanding of geometric images and corpora. Furthermore, we introduce geometry-language alignment, an effective pre-training paradigm that bridges the modality gap between unimodal geometric experts. We propose a Generator-And-Sampler Transformer (GS-Former) to generate discriminative queries and eliminate uninformative representations from unevenly distributed geometric signals. Finally, GeoX benefits from visual instruction tuning, empowering it to take geometric images and questions as input and generate verifiable solutions. Experiments show that GeoX outperforms both generalists and geometric specialists on publicly recognized benchmarks, such as GeoQA, UniGeo, Geometry3K, and PGPS9k. Our data and code will be released soon to accelerate future research on automatic GPS.

## 2566. Feedback Schrödinger Bridge Matching

链接：https://iclr.cc/virtual/2025/poster/28600 abstract： Recent advancements in diffusion bridges for distribution transport problems have heavily relied on matching frameworks, yet existing methods often face a trade-off between scalability and access to optimal pairings during training. Fully unsupervised methods make minimal assumptions but incur high computational costs, limiting their practicality. On the other hand, imposing full supervision of the matching process with optimal pairings improves scalability, however, it can be infeasible in most applications.To strike a balance between scalability and minimal supervision, we introduce Feedback Schrödinger Bridge Matching (FSBM), a novel semi-supervised matching framework that incorporates a small portion ($<8$% of the entire dataset) of pre-aligned pairs as state feedback to guide the transport map of non-coupled samples, thereby significantly improving efficiency. This is achieved by formulating a static Entropic Optimal Transport (EOT) problem with an additional term capturing the semi-supervised guidance. The generalized EOT objective is then recast into a dynamic formulation to leverage the scalability of matching frameworks. Extensive experiments demonstrate that FSBM accelerates training and enhances generalization by leveraging coupled pairs' guidance, opening new avenues for training matching frameworks with partially aligned datasets.

## 2567. Drama: Mamba-Enabled Model-Based Reinforcement Learning Is Sample and Parameter Efficient

链接：https://iclr.cc/virtual/2025/poster/30818 abstract： Model-based reinforcement learning (RL) offers a solution to the data inefficiency that plagues most model-free RL algorithms. However, learning a robust world model often requires complex and deep architectures, which are computationally expensive and challenging to train. Within the world model, sequence models play a critical role in accurate predictions, and various architectures have been explored, each with its own challenges. Currently, recurrent neural network (RNN)-based world models struggle with vanishing gradients and capturing long-term dependencies. Transformers, on the other hand, suffer from the quadratic memory and computational complexity of self-attention mechanisms, scaling as $O(n^2)$, where $n$ is the sequence length.To address these challenges, we propose a state space model (SSM)-based world model, Drama, specifically leveraging Mamba, that achieves $O(n)$ memory and computational complexity while effectively capturing long-term dependencies and enabling efficient training with longer sequences. We also introduce a novel sampling method to mitigate the suboptimality caused by an incorrect world model in the early training stages. Combining these techniques, Drama achieves a normalised score on the Atari100k benchmark that is competitive with other state-of-the-art (SOTA) model-based RL algorithms, using only a 7 million-parameter world model. Drama is accessible and trainable on off-the-shelf hardware, such as a standard laptop. Our code is available at https://github.com/realwenlongwang/Drama.git.

## 2568. Deep Distributed Optimization for Large-Scale Quadratic Programming

链接：https://iclr.cc/virtual/2025/poster/28726 abstract： Quadratic programming (QP) forms a crucial foundation in optimization, appearing in a broad spectrum of domains and serving as the basis for more advanced algorithms. Consequently, as the scale and complexity of modern applications continue to grow, the development of efficient and reliable QP algorithms becomes increasingly vital. In this context, this paper introduces a novel deep learning-aided distributed optimization architecture designed for tackling large-scale QP problems. First, we combine the state-of-the-art Operator Splitting QP (OSQP) method with a consensus approach to derive DistributedQP, a new method tailored for network-structured problems, with convergence guarantees to optimality. Subsequently, we unfold this optimizer into a deep learning framework, leading to DeepDistributedQP, which leverages learned policies to accelerate reaching to desired accuracy within a restricted amount of iterations. Our approach is also theoretically grounded through Probably Approximately Correct (PAC)-Bayes theory, providing generalization bounds on the expected optimality gap for unseen problems. The proposed framework, as well as its centralized version DeepQP, significantly outperform their standard optimization counterparts on a variety of tasks such as randomly generated problems, optimal control, linear regression, transportation networks and others. Notably, DeepDistributedQP demonstrates strong generalization by training on small problems and scaling to solve much larger ones (up to 50K variables and 150K constraints) using the same policy. Moreover, it achieves orders-of-magnitude improvements in wall-clock time compared to OSQP. The certifiable performance guarantees of our approach are also demonstrated, ensuring higher-quality solutions over traditional optimizers.

## 2569. Field-DiT: Diffusion Transformer on Unified Video, 3D, and Game Field Generation

链接：https://iclr.cc/virtual/2025/poster/27874 abstract： The probabilistic field models the distribution of continuous functions defined over metric spaces. While these models hold great potential for unifying data generation across various modalities, including images, videos, and 3D geometry, they still struggle with long-context generation beyond simple examples. This limitation can be attributed to their MLP architecture, which lacks sufficient inductive bias to capture global structures through uniform sampling. To address this, we propose a new and simple model that incorporates a view-wise sampling algorithm to focus on local structure learning, along with autoregressive generation to preserve global geometry. It adapts cross-modality conditions, such as text prompts for text-to-video generation, camera poses for 3D view generation, and control actions for game generation. Experimental results across various modalities demonstrate the effectiveness of our model, with its 675M parameter size, and highlight its potential as a foundational framework for scalable, architecture-unified visual content generation for different modalities with different weights. Our project page can be found at https://kfmei.com/Field-DiT/.

## 2570. Safety Layers in Aligned Large Language Models: The Key to LLM Security

链接：https://iclr.cc/virtual/2025/poster/28578 abstract： Aligned LLMs are secure, capable of recognizing and refusing to answer malicious questions. However, the role of internal parameters in maintaining such security is not well understood yet, further these models can be vulnerable to security degradation when subjected to fine-tuning attacks. To address these challenges, our work uncovers the mechanism behind security in aligned LLMs at the parameter level, identifying a small set of contiguous layers in the middle of the model that are crucial for distinguishing malicious queries from normal ones, referred to as "safety layers". We first confirm the existence of these safety layers by analyzing variations in input vectors within the model's internal layers. Additionally, we leverage the over-rejection phenomenon and parameters scaling analysis to precisely locate the safety layers. Building on these findings, we propose a novel fine-tuning approach, Safely Partial-Parameter Fine-Tuning (SPPFT), that fixes the gradient of the safety layers during fine-tuning to address the security degradation. Our experiments demonstrate that the proposed approach can significantly preserve LLM security while maintaining performance and reducing computational resources compared to full fine-tuning.

## 2571. PIN: Prolate Spheroidal Wave Function-based Implicit Neural Representations

链接：https://iclr.cc/virtual/2025/poster/30391 abstract： Implicit Neural Representations (INRs) provide a continuous mapping between the coordinates of a signal and the corresponding values. As the performance of INRs heavily depends on the choice of nonlinear-activation functions, there has been a significant focus on encoding explicit signals within INRs using diverse activation functions. Despite recent advancements, existing INRs often encounter significant challenges, particularly at fine scales where they often introduce noise-like artifacts over smoother areas compromising the quality of the output. Moreover, they frequently struggle to generalize to unseen coordinates. These drawbacks highlight a critical area for further research and development to enhance the robustness and applicability of INRs across diverse scenarios. To address this challenge, we introduce the Prolate Spheroidal Wave Function-based Implicit Neural Representations (PIN), which exploits the optimal space-frequency domain concentration of Prolate Spheroidal Wave Functions (PSWFs) as the nonlinear mechanism in INRs. Our experimental results reveal that PIN excels not only in representing images and 3D shapes but also significantly outperforms existing methods in various vision tasks that require INR generalization, including image inpainting, novel view synthesis, edge detection, and image denoising.

## 2572. Learning Harmonized Representations for Speculative Sampling

链接：https://iclr.cc/virtual/2025/poster/29560 abstract： Speculative sampling is a promising approach to accelerate the decoding stage for Large Language Models (LLMs). Recent advancements that leverage target LLM's contextual information, such as hidden states and KV cache, have shown significant practical improvements. However, these approaches suffer from inconsistent context between training and decoding. We also observe another discrepancy between the training and decoding objectives in existing speculative sampling methods. In this work, we propose a solution named HArmonized Speculative Sampling (HASS) that learns harmonized representations to address these issues. HASS accelerates the decoding stage without adding inference overhead through harmonized objective distillation and harmonized context alignment. Experiments on four LLaMA models demonstrate that HASS achieves 2.81x-4.05x wall-clock time speedup ratio averaging across three datasets, surpassing EAGLE-2 by 8%-20%. The code is available at https://github.com/HArmonizedSS/HASS.

## 2573. High-Precision Dichotomous Image Segmentation via Probing Diffusion Capacity

链接：https://iclr.cc/virtual/2025/poster/27897 abstract： In the realm of high-resolution (HR), fine-grained image segmentation, the primary challenge is balancing broad contextual awareness with the precision required for detailed object delineation, capturing intricate details and the finest edges of objects. Diffusion models, trained on vast datasets comprising billions of image-text pairs, such as SD V2.1, have revolutionized text-to-image synthesis by delivering exceptional quality, fine detail resolution, and strong contextual awareness, making them an attractive solution for high-resolution image segmentation. To this end, we propose DiffDIS, a diffusion-driven segmentation model that taps into the potential of the pre-trained U-Net within diffusion models, specifically designed for high-resolution, fine-grained object segmentation. By leveraging the robust

generalization capabilities and rich, versatile image representation prior of the SD models, coupled with a task-specific stable one-step denoising approach, we significantly reduce the inference time while preserving high-fidelity, detailed generation. Additionally, we introduce an auxiliary edge generation task to not only enhance the preservation of fine details of the object boundaries, but reconcile the probabilistic nature of diffusion with the deterministic demands of segmentation. With these refined strategies in place, DiffDIS serves as a rapid object mask generation model, specifically optimized for generating detailed binary maps at high resolutions, while demonstrating impressive accuracy and swift processing. Experiments on the DIS5K dataset demonstrate the superiority of DiffDIS, achieving state-of-the-art results through a streamlined inference process. The source code will be publicly available at \href{https://github.com/qianyu-dlut/DiffDIS}{DiffDIS}.

## 2574. PEAR: Primitive Enabled Adaptive Relabeling for Boosting Hierarchical Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/31238 abstract： Hierarchical reinforcement learning (HRL) has the potential to solve complex long horizon tasks using temporal abstraction and increased exploration. However, hierarchical agents are difficult to train due to inherent non-stationarity. We present primitive enabled adaptive relabeling (PEAR), a two-phase approach where we first perform adaptive relabeling on a few expert demonstrations to generate efficient subgoal supervision, and then jointly optimize HRL agents by employing reinforcement learning (RL) and imitation learning (IL). We perform theoretical analysis to bound the sub-optimality of our approach and derive a joint optimization framework using RL and IL. Since PEAR utilizes only a few expert demonstrations and considers minimal limiting assumptions on the task structure, it can be easily integrated with typical off-policy \RL algorithms to produce a practical HRL approach. We perform extensive experiments on challenging environments and show that PEAR is able to outperform various hierarchical and non-hierarchical baselines and achieve upto 80% success rates in complex sparse robotic control tasks where other baselines typically fail to show significant progress. We also perform ablations to thoroughly analyze the importance of our various design choices. Finally, we perform real world robotic experiments on complex tasks and demonstrate that PEAR consistently outperforms the baselines.

## 2575. REEF: Representation Encoding Fingerprints for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29580 abstract： Protecting the intellectual property of open-source Large Language Models (LLMs) is very important, because training LLMs costs extensive computational resources and data. Therefore, model owners and third parties need to identify whether a suspect model is a subsequent development of the victim model. To this end, we propose a training-free REEF to identify the relationship between the suspect and victim models from the perspective of LLMs' feature representations. Specifically, REEF computes and compares the centered kernel alignment similarity between the representations of a suspect model and a victim model on the same samples. This training-free REEF does not impair the model's general capabilities and is robust to sequential fine-tuning, pruning, model merging, and permutations. In this way, REEF provides a simple and effective way for third parties and models' owners to protect LLMs' intellectual property together. Our code is publicly accessible at https://github.com/AI45Lab/REEF.

## 2576. From Tokens to Lattices: Emergent Lattice Structures in Language Models

链接：https://iclr.cc/virtual/2025/poster/28457 abstract： Pretrained masked language models (MLMs) have demonstrated an impressive capability to comprehend and encode conceptual knowledge, revealing a lattice structure among concepts. This raises a critical question: how does this conceptualization emerge from MLM pretraining? In this paper, we explore this problem from the perspective of Formal Concept Analysis (FCA), a mathematical framework that derives concept lattices from the observations of object-attribute relationships. We show that the MLM's objective implicitly learns a formal context that describes objects, attributes, and their dependencies, which enables the reconstruction of a concept lattice through FCA. We propose a novel framework for concept lattice construction from pretrained MLMs and investigate the origin of the inductive biases of MLMs in lattice structure learning. Our framework differs from previous work because it does not rely on human-defined concepts and allows for discovering "latent" concepts that extend beyond human definitions. We create three datasets for evaluation, and the empirical results verify our hypothesis.

## 2577. The Computational Complexity of Positive Non-Clashing Teaching in Graphs

链接：https://iclr.cc/virtual/2025/poster/30095 abstract： We study the classical and parameterized complexity of computing the positive non-clashing teaching dimension of a set of concepts, that is, the smallest number of examples per concept required to successfully teach an intelligent learner under the considered, previously established model. For any class of concepts, it is known that this problem can be effortlessly transferred to the setting of balls in a graph $G$. We establish (1) the NP-hardness of the problem even when restricted to instances with positive non-clashing teaching dimension $k=2$ and where all balls in the graph are present, (2) near-tight running time upper and lower bounds for the problem on general graphs, (3) fixed-parameter tractability when parameterized by the vertex integrity of $G$, and (4) a lower bound excluding fixed-parameter tractability when parameterized by the feedback vertex number and pathwidth of $G$, even when combined with $k$.Our results provide a nearly complete understanding of the complexity landscape of computing the positive non-clashing teaching dimension and answer

open questions from the literature.

## 2578. Smaller, Weaker, Yet Better: Training LLM Reasoners via Compute-Optimal Sampling

链接：https://iclr.cc/virtual/2025/poster/31080 abstract： Training on high-quality synthetic data from strong language models (LMs) is a common strategy to improve the reasoning performance of LMs. In this work, we revisit whether this strategy is compute-optimal under a fixed inference budget (e.g., FLOPs). To do so, we investigate the trade-offs between generating synthetic data using a stronger but more expensive (SE) model versus a weaker but cheaper (WC) model. We evaluate the generated data across three key metrics: coverage, diversity, and false positive rate, and show that the data from WC models may have higher coverage and diversity, but also exhibit higher false positive rates. We then finetune LMs on data from SE and WC models in different settings: knowledge distillation, self-improvement, and a novel weak-to-strong improvement setup where a weaker LM teaches reasoning to a stronger LM. Our findings reveal that models finetuned on WC-generated data consistently outperform those trained on SE-generated data across multiple benchmarks and multiple choices of WC and SE models. These results challenge the prevailing practice of relying on SE models for synthetic data generation, suggesting that WC may be the compute-optimal approach for training advanced LM reasoners.

## 2579. Accelerating Diffusion Transformers with Token-wise Feature Caching

链接：https://iclr.cc/virtual/2025/poster/27718 abstract： Diffusion transformers have shown significant effectiveness in both image and video synthesis at the expense of huge computation costs. To address this problem, feature caching methods have been introduced to accelerate diffusion transformers by caching the features in previous timesteps and reusing them in the following timesteps. However, previous caching methods ignore that different tokens exhibit different sensitivities to feature caching, and feature caching on some tokens may lead to 10$\times$ more destruction to the overall generation quality compared with other tokens. In this paper, we introduce token-wise feature caching, allowing us to adaptively select the most suitable tokens for caching, and further enable us to apply different caching ratios to neural layers in different types and depths. Extensive experiments on PixArt-alpha, OpenSora, and DiT demonstrate our effectiveness in both image and video generation with no requirements for training. For instance, 2.36$\times$ and 1.93$\times$ acceleration are achieved on OpenSora and PixArt-$\alpha$ with almost no drop in generation quality. Codes have been released in the supplementary material and Github.

## 2580. Revisiting Prefix-tuning: Statistical Benefits of Reparameterization among Prompts

链接：https://iclr.cc/virtual/2025/poster/29676 abstract： Prompt-based techniques, such as prompt-tuning and prefix-tuning, have gained prominence for their efficiency in fine-tuning large pre-trained models. Despite their widespread adoption, the theoretical foundations of these methods remain limited. For instance, in prefix-tuning, we observe that a key factor in achieving performance parity with full fine-tuning lies in the reparameterization strategy. However, the theoretical principles underpinning the effectiveness of this approach have yet to be thoroughly examined. Our study demonstrates that reparameterization is not merely an engineering trick but is grounded in deep theoretical foundations. Specifically, we show that the reparameterization strategy implicitly encodes a shared structure between prefix key and value vectors. Building on recent insights into the connection between prefix-tuning and mixture of experts models, we further illustrate that this shared structure significantly improves sample efficiency in parameter estimation compared to non-shared alternatives. The effectiveness of prefix-tuning across diverse tasks is empirically confirmed to be enhanced by the shared structure, through extensive experiments in both visual and language domains. Additionally, we uncover similar structural benefits in prompt-tuning, offering new perspectives on its success. Our findings provide theoretical and empirical contributions, advancing the understanding of prompt-based methods and their underlying mechanisms.

## 2581. Extendable and Iterative Structure Learning Strategy for Bayesian Networks

链接：https://iclr.cc/virtual/2025/poster/31054 abstract： Learning the structure of Bayesian networks is a fundamental yet computationally intensive task, especially as the number of variables grows. Traditional algorithms require retraining from scratch when new variables are introduced, making them impractical for dynamic or large-scale applications. In this paper, we propose an extendable structure learning strategy that efficiently incorporates a new variable $Y$ into an existing Bayesian network graph $\mathcal{G}$ over variables $\mathcal{X}$, resulting in an updated P-map graph $\bar{\mathcal{G}}$ on $\bar{\mathcal{X}} = \mathcal{X} \cup {Y}$. By leveraging the information encoded in $\mathcal{G}$, our method significantly reduces computational overhead compared to learning $\bar{\mathcal{G}}$ from scratch. Empirical evaluations demonstrate runtime reductions of up to 1300x without compromising accuracy. Building on this approach, we introduce a novel iterative paradigm for structure learning over $\mathcal{X}$. Starting with a small subset $\mathcal{U} \subset \mathcal{X}$, we iteratively add the remaining variables using our extendable algorithms to construct a P-map graph over the full set. This method offers runtime advantages comparable to common algorithms while maintaining similar accuracy. Our contributions provide a scalable solution for Bayesian network structure learning, enabling efficient model updates in real-time and high-dimensional settings.

## 2582. Gnothi Seauton: Empowering Faithful Self-Interpretability in Black-Box

# Transformers

链接：https://iclr.cc/virtual/2025/poster/29444 abstract： The debate between self-interpretable models and post-hoc explanations for black-box models is central to Explainable AI (XAI). Self-interpretable models, such as concept-based networks, offer insights by connecting decisions to human-understandable concepts but often struggle with performance and scalability. Conversely, post-hoc methods like Shapley values, while theoretically robust, are computationally expensive and resource-intensive. To bridge the gap between these two lines of research, we propose a novel method that combines their strengths, providing theoretically guaranteed self-interpretability for black-box models without compromising prediction accuracy. Specifically, we introduce a parameter-efficient pipeline, AutoGnothi, which integrates a small side network into the black-box model, allowing it to generate Shapley value explanations without changing the original network parameters. This side-tuning approach significantly reduces memory, training, and inference costs, outperforming traditional parameter-efficient methods, where full fine-tuning serves as the optimal baseline. AutoGnothi enables the black-box model to predict and explain its predictions with minimal overhead. Extensive experiments show that AutoGnothi offers accurate explanations for both vision and language tasks, delivering superior computational efficiency with comparable interpretability.

# 2583. Efficient Reinforcement Learning with Large Language Model Priors

链接：https://iclr.cc/virtual/2025/poster/28953 abstract： In sequential decision-making (SDM) tasks, methods like reinforcement learning (RL) and heuristic search have made notable advances in specific cases. However, they often require extensive exploration and face challenges in generalizing across diverse environments due to their limited grasp of the underlying decision dynamics. In contrast, large language models (LLMs) have recently emerged as powerful general-purpose tools, due to their capacity to maintain vast amounts of domain-specific knowledge. To harness this rich prior knowledge for efficiently solving complex SDM tasks, we propose treating LLMs as prior action distributions and integrating them into RL frameworks through Bayesian inference methods, making use of variational inference and direct posterior sampling. The proposed approaches facilitate the seamless incorporation of fixed LLM priors into both policy-based and value-based RL frameworks. Our experiments show that incorporating LLM-based action priors significantly reduces exploration and optimization complexity, substantially improving sample efficiency compared to traditional RL techniques, e.g., using LLM priors decreases the number of required samples by over 90\% in offline learning scenarios.

# 2584. KinFormer: Generalizable Dynamical Symbolic Regression for Catalytic Organic Reaction Kinetics

链接：https://iclr.cc/virtual/2025/poster/28399 abstract： Modeling kinetic equations is essential for understanding the mechanisms of chemical reactions, yet a complex and time-consuming task. Kinetic equation prediction is formulated as a problem of dynamical symbolic regression (DSR) subject to physical chemistry constraints. Deep learning (DL) holds the potential to capture reaction patterns and predict kinetic equations from data of chemical species, effectively avoiding empirical bias and improving efficiency compared with traditional analytical methods. Despite numerous studies focusing on DSR and the introduction of Transformers to predict ordinary differential equations, the corresponding models lack generalization abilities across diverse categories of reactions. In this study, we propose KinFormer, a generalizable kinetic equation prediction model. KinFormer utilizes a conditional Transformer to model DSR under physical constraints and employs Monte Carlo Tree Search to apply the model to new types of reactions. Experimental results on 20 types of organic reactions demonstrate that KinFormer not only outperforms classical baselines, but also exceeds Transformer baselines in out-of-domain evaluations, thereby proving its generalization ability.

# 2585. Transformers Provably Solve Parity Efficiently with Chain of Thought

链接：https://iclr.cc/virtual/2025/poster/28433 abstract： This work provides the first theoretical analysis of training transformers to solve complex problems by recursively generating intermediate states, analogous to fine-tuning for chain-of-thought (CoT) reasoning. We consider training a one-layer transformer to solve the fundamental $k$-parity problem, extending the work on RNNs by \citet{Wies23}. We establish three key results: (1) any finite-precision gradient-based algorithm, without intermediate supervision, requires substantial iterations to solve parity with finite samples. (2) In contrast, when intermediate parities are incorporated into the loss function, our model can learn parity in one gradient update when aided by \emph{teacher forcing}, where ground-truth labels of the reasoning chain are provided at each generation step. (3) Even without teacher forcing, where the model must generate CoT chains end-to-end, parity can be learned efficiently if augmented data is employed to internally verify the soundness of intermediate steps. Our findings, supported by numerical experiments, show that task decomposition and stepwise reasoning naturally arise from optimizing transformers with CoT; moreover, self-consistency checking can improve multi-step reasoning ability, aligning with empirical studies of CoT.

# 2586. Rethinking Classifier Re-Training in Long-Tailed Recognition: Label Over-Smooth Can Balance

链接：https://iclr.cc/virtual/2025/poster/29813 abstract： In the field of long-tailed recognition, the Decoupled Training paradigm has shown exceptional promise by dividing training into two stages: representation learning and classifier re-training. While previous work has tried to improve both stages simultaneously, this complicates isolating the effect of classifier re-training.

Recent studies reveal that simple regularization can produce strong feature representations, highlighting the need to reassess classifier re-training methods. In this study, we revisit classifier re-training methods based on a unified feature representation and re-evaluate their performances. We propose two new metrics, Logits Magnitude and Regularized Standard Deviation, to compare the differences and similarities between various methods. Using these two newly proposed metrics, we demonstrate that when the Logits Magnitude across classes is nearly balanced, further reducing its overall value can effectively decrease errors and disturbances during training, leading to better model performance. Based on our analysis using these metrics, we observe that adjusting the logits could improve model performance, leading us to develop a simple label over-smoothing approach to adjust the logits without requiring prior knowledge of class distribution.This method softens the original one-hot labels by assigning a probability slightly higher than $\frac{1}{K}$ to the true class and slightly lower than $\frac{1}{K}$ to the other classes, where $K$ is the number of classes.Our method achieves state-of-the-art performance on various imbalanced datasets, including CIFAR100-LT, ImageNet-LT, and iNaturalist2018.

## 2587. Pedestrian Motion Reconstruction: A Large-scale Benchmark via Mixed Reality Rendering with Multiple Perspectives and Modalities

链接：https://iclr.cc/virtual/2025/poster/29263 abstract： Reconstructing pedestrian motion from dynamic sensors, with a focus on pedestrian intention, is crucial for advancing autonomous driving safety. However, this task is challenging due to data limitations arising from technical complexities, safety, and cost concerns. We introduce the Pedestrian Motion Reconstruction (PMR) dataset, which focuses on pedestrian intention to reconstruct behavior using multiple perspectives and modalities. PMR is developed from a mixed reality platform that combines real-world realism with the extensive, accurate labels of simulations, thereby reducing costs and risks. It captures the intricate dynamics of pedestrian interactions with objects and vehicles, using different modalities for a comprehensive understanding of human-vehicle interaction. Analyses show that PMR can naturally exhibit pedestrian intent and simulate extreme cases. PMR features a vast collection of data from 54 subjects interacting across 12 urban settings with 7 objects, encompassing 12,138 sequences with diverse weather conditions and vehicle speeds. This data provides a rich foundation for modeling pedestrian intent through multi-view and multi-modal insights. We also conduct comprehensive benchmark assessments across different modalities to thoroughly evaluate pedestrian motion reconstruction methods.

## 2588. IDArb: Intrinsic Decomposition for Arbitrary Number of Input Views and Illuminations

链接：https://iclr.cc/virtual/2025/poster/27948 abstract： Capturing geometric and material information from images remains a fundamental challenge in computer vision and graphics. Traditional optimization-based methods often require hours of computational time to reconstruct geometry, material properties, and environmental lighting from dense multi-view inputs, while still struggling with inherent ambiguities between lighting and material. On the other hand, learning-based approaches leverage rich material priors from existing 3D object datasets but face challenges with maintaining multi-view consistency.In this paper, we introduce IDArb, a diffusion-based model designed to perform intrinsic decomposition on an arbitrary number of images under varying illuminations. Our method achieves highly accurate and multi-view consistent estimation on surface normals and material properties. This is made possible through a novel cross-view, cross-domain attention module and an illumination-augmented, view-adaptive training strategy. Additionally, we introduce ARB-Objaverse, a new dataset that provides large-scale multi-view intrinsic data and renderings under diverse lighting conditions, supporting robust training.Extensive experiments demonstrate that IDArb outperforms state-of-the-art methods both qualitatively and quantitatively. Moreover, our approach facilitates a range of downstream tasks, including single-image relighting, photometric stereo, and 3D reconstruction, highlighting its broad applicability in realistic 3D content creation.Project website: https://lizb6626.github.io/IDArb/.

## 2589. SimpleTM: A Simple Baseline for Multivariate Time Series Forecasting

链接：https://iclr.cc/virtual/2025/poster/28373 abstract： The versatility of large Transformer-based models has led to many efforts focused on adaptations to other modalities, including time-series data. For instance, one could start from a pre-trained checkpoint of a large language model and attach adapters to recast the new modality (e.g., time-series) as ``language''. Alternatively, one can use a suitably large Transformer-based model, and make some modifications for time-series data. These ideas offer good performance across available benchmarks. But temporal data are quite heterogeneous (e.g., wearable sensors, physiological measurements in healthcare), and unlike text/image corpus, much of it is not publicly available. So, these models need a fair bit of domain-specific fine-tuning to achieve good performance -- this is often expensive or difficult with limited resources. In this paper, we study and characterize the performance profile of a non-generalist approach: our SimpleTM model is specialized for multivariate time-series forecasting. By simple, we mean that the model is lightweight. It is restricted to tokenization based on textbook signal processing ideas (shown to be effective in vision) which are then allowed to attend/interact: via self-attention but also via ways that are a bit more general than dot-product attention, accomplished via basic geometric algebra operations. We show that even a single- or two-layer model gives results that are competitive with much bigger models, including large transformer-based architectures,on most benchmarks commonly reported in the literature.

## 2590. CLIPure: Purification in Latent Space via CLIP for Adversarially Robust Zero-Shot Classification

链接：https://iclr.cc/virtual/2025/poster/29543 abstract： In this paper, we aim to build an adversarially robust zero-shot image classifier that can accurately and efficiently classify unseen examples while defending against unforeseen adversarial attacks, addressing critical challenges in real-world safety-sensitive scenarios. To achieve this, we focus on two key challenges: zero-shot classification and defense against unforeseen attacks. We ground our work on CLIP, a vision-language pre-trained model to perform zero-shot classification. To defend against unforeseen attacks, we adopt a purification approach, as it is independent of specific attack types. We then define a purification risk as the KL divergence between the joint distributions of the purification and attack process. The derived lower bound of purification risk inspires us to explore purification in CLIP's multi-modal latent space. We propose a CLIP-based purification method called CLIPure, which has two variants: CLIPure-Diff, which models image likelihood with a generative process of its latent vector, and CLIPure-Cos, which models the likelihood based on the similarity between embeddings of the image and a blank template. As far as we know, CLIPure is the first purification method in latent space and CLIPure-Cos is the first purification method not relying on generative models, substantially improving defense efficiency. Extensive experimental results show that the robustness achieved by CLIPure is within a small gap of clean accuracy, outperforming SOTA robustness by a large margin, e.g., from 71.7\% to 91.1\% on CIFAR10, from 59.6\% to 72.6\% on ImageNet, and 108\% relative improvements of average robustness on the 13 datasets over previous SOTA, with only 14\% extra inference cost and no additional training.

## 2591. Models trained with unnormalized density functions: A need for a course correction

链接：https://iclr.cc/virtual/2025/poster/31333 abstract： Training a generative model with energy or unnormalized density functions is considered an important problem for physical systems such as molecules. This provides a path to train generative models to sample from the much desired Boltzmann distribution in situations of data scarcity. As of late, several generative frameworks have been proposed to target this problem. However, as we show in the following blog post, these methods have not been benchmarked sufficiently well against traditional Markov Chain Monte Carlo (MCMC) methods that are used to sample from energy functions. We take the example of two recent methods (IDEM and IEFM) and show that MCMC outperforms both methods in terms of number of energy evaluations and wall clock time on established baselines. With this, we suggest a “course correction” on the benchmarking of these models and comment on the utility and potential of generative models on these tasks.

## 2592. Composing Unbalanced Flows for Flexible Docking and Relaxation

链接：https://iclr.cc/virtual/2025/poster/28828 abstract： Diffusion models have emerged as a successful approach for molecular docking, but they often cannot model protein flexibility or generate nonphysical poses. We argue that both these challenges can be tackled by framing the problem as a transport between distributions. Still, existing paradigms lack the flexibility to define effective maps between such complex distributions. To address this limitation we propose Unbalanced Flow Matching, a generalization of Flow Matching (FM) that allows trading off sample efficiency with approximation accuracy and enables more accurate transport. Empirically, we apply Unbalanced FM on flexible docking and structure relaxation, demonstrating our ability to model protein flexibility and generate energetically favorable poses. On the PDBBind docking benchmark, our method FlexDock improves the docking performance while increasing the proportion of energetically favorable poses from 30% to 73%.

## 2593. REMEDY: Recipe Merging Dynamics in Large Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/28692 abstract： Model merging has emerged as a powerful technique for combining task-specific vision models into a unified and multi-functional model. Previous methods represented by task arithmetic, have demonstrated effectiveness and scalability in this domain. When large vision-language models (LVLMs) arise with model size scaling up, this design becomes challenging to fuse different instruction-tuned LVLMs for generalization enhancement. The large scale and multi-modal nature of LVLMs present unique obstacles, including constructing reusable and modular components to accommodate the multi-component architecture of LVLMs and the requirement for dynamic fusion based on multi-modal input tokens. To address these challenges, we propose the \textbf{RE}cipe \textbf{ME}rging \textbf{DY}namics (REMEDY) method, a scalable and flexible paradigm for model merging in LVLMs. We first define reusable modules termed \textit{recipes} including the projector and shallow LLM layers, enhancing visual-language understanding. Then, we introduce a modality-aware allocator dynamically generates weights in a one-shot manner based on input relevance to existing recipes, enabling efficient cross-modal knowledge integration. REMEDY thus offers an adaptive solution for LVLMs to tackle both seen (i.e., multi-task learning) and unseen (i.e., zero-shot generalization) tasks. Experimental results demonstrate that our method consistently improves performance on both seen and unseen tasks, underscoring the effectiveness of REMEDY in diverse multi-modal scenarios.

## 2594. Generalized Behavior Learning from Diverse Demonstrations

链接：https://iclr.cc/virtual/2025/poster/29708 abstract： Diverse behavior policies are valuable in domains requiring quick test-time adaptation or personalized human-robot interaction. Human demonstrations provide rich information regarding task objectives and factors that govern individual behavior variations, which can be used to characterize \textit{useful} diversity and learn diverse performant policies.However, we show that prior work that builds naive representations of demonstration heterogeneity fails in generating successful novel behaviors that generalize over behavior factors.We propose Guided Strategy Discovery (GSD), which introduces a novel diversity formulation based on a learned task-relevance measure that prioritizes behaviors exploring modeled latent factors.We empirically validate across three continuous control benchmarks for generalizing to in-distribution (interpolation) and out-of-distribution (extrapolation) factors that GSD outperforms baselines in novel behavior

discovery by $\sim$21\%.Finally, we demonstrate that GSD can generalize striking behaviors for table tennis in a virtual testbed while leveraging human demonstrations collected in the real world.Code is available at https://github.com/CORE-Robotics-Lab/GSD.

## 2595. Semi-Supervised CLIP Adaptation by Enforcing Semantic and Trapezoidal Consistency

链接：https://iclr.cc/virtual/2025/poster/30720 abstract： Vision-language pre-training models, such as CLIP, have demonstrated strong capability in rapidly adapting to downstream tasks through fine-tuning, and have been widely applied across various tasks. However, when the downstream tasks are constrained by limited image-text paired data, CLIP struggles to effectively address the domain gap between the pre-training and the target tasks. To address this limitation, we propose a novel semi-supervised CLIP training method coined SemiCLIP that leverages a small amount of image-text pairs alongside a large volume of images without text descriptions to enhance CLIP's cross-modal alignment. To effectively utilize unlabeled images, we introduce semantic concept mining to improve task-specific visual representations by matching images with relevant concepts mined from labeled data. Leveraging matched semantic concepts, we construct learnable surrogate captions for unlabeled images and optimize a trapezoidal consistency to regulate the geometric structure of image-text pairs in the representation space. Experimental results demonstrate that our approach significantly improves the adaptability of CLIP in target tasks with limited labeled data, achieving gains ranging from 1.72\% -- 6.58\% for zero-shot classification accuracy and 2.32\% -- 3.23\% for image-text retrieval performance on standard benchmarks. The source code is available at https://github.com/Gank0078/SemiCLIP.

## 2596. Holographic Node Representations: Pre-training Task-Agnostic Node Embeddings

链接：https://iclr.cc/virtual/2025/poster/28064 abstract： Large general purpose pre-trained models have revolutionized computer vision and natural language understanding. However, the development of general purpose pre-trained Graph Neural Networks (GNNs) lags behind other domains due to the lack of suitable generalist node representations. Existing GNN architectures are often tailored to specific task orders, such as node-level, link-level, or higher-order tasks, because different tasks require distinct permutation symmetries, which are difficult to reconcile within a single model. In this paper, we propose holographic node representations, a new blueprint for node representations capable of solving tasks of any order. Holographic node representations have two key components: (1) a task-agnostic expansion map, which produces highly expressive, high-dimensional embeddings, free from node-permutation symmetries, to be fed into (2) a reduction map that carefully reintroduces the relevant permutation symmetries to produce low-dimensional, task-specific embeddings. We show that well-constructed expansion maps enable simple and efficient reduction maps, which can be adapted for any task order. Empirical results show that holographic node representations can be effectively pre-trained and reused across tasks of varying orders, yielding up to 100% relative performance improvement, including in cases where prior methods fail entirely.

## 2597. Realistic Evaluation of Deep Partial-Label Learning Algorithms

链接：https://iclr.cc/virtual/2025/poster/30321 abstract： Partial-label learning (PLL) is a weakly supervised learning problem in whicheach example is associated with multiple candidate labels and only one is thetrue label. In recent years, many deep PLL algorithms have been developed toimprove model performance. However, we find that some early developedalgorithms are often underestimated and can outperform many later algorithmswith complicated designs. In this paper, we delve into the empiricalperspective of PLL and identify several critical but previously overlookedissues. First, model selection for PLL is non-trivial, but has never beensystematically studied. Second, the experimental settings are highlyinconsistent, making it difficult to evaluate the effectiveness of thealgorithms. Third, there is a lack of real-world image datasets that can becompatible with modern network architectures. Based on these findings, wepropose PLENCH, the first Partial-Label learning bENCHmark to systematicallycompare state-of-the-art deep PLL algorithms. We investigate the modelselection problem for PLL for the first time, and propose novel model selectioncriteria with theoretical guarantees. We also create Partial-Label CIFAR-10(PLCIFAR10), an image dataset of human-annotated partial labels collected fromAmazon Mechanical Turk, to provide a testbed for evaluating the performance ofPLL algorithms in more realistic scenarios. Researchers can quickly andconveniently perform a comprehensive and fair evaluation and verify theeffectiveness of newly developed algorithms based on PLENCH. We hope thatPLENCH will facilitate standardized, fair, and practical evaluation of PLLalgorithms in the future.

## 2598. Mitigating Modality Prior-Induced Hallucinations in Multimodal Large Language Models via Deciphering Attention Causality

链接：https://iclr.cc/virtual/2025/poster/30629 abstract： Multimodal Large Language Models (MLLMs) have emerged as a central focus in both industry and academia, but often suffer from biases introduced by visual and language priors, which can lead to multimodal hallucination. These biases arise from the visual encoder and the Large Language Model (LLM) backbone, affecting the attention mechanism responsible for aligning multimodal inputs. Existing decoding-based mitigation methods focus on statistical correlations and overlook the causal relationships between attention mechanisms and model output, limiting their effectiveness in addressing these biases. To tackle this issue, we propose a causal inference framework termed CausalMM that applies structural causal modeling to MLLMs, treating modality priors as a confounder between attention mechanisms and

output. Specifically, by employing backdoor adjustment and counterfactual reasoning at both the visual and language attention levels, our method mitigates the negative effects of modality priors and enhances the alignment of MLLM's inputs and outputs, with a maximum score improvement of 65.3% on 6 VLind-Bench indicators and 164 points on MME Benchmark compared to conventional methods. Extensive experiments validate the effectiveness of our approach while being a plug-and-play solution. Our code is available at: https://github.com/The-Martyr/CausalMM.

## 2599. PostEdit: Posterior Sampling for Efficient Zero-Shot Image Editing

链接：https://iclr.cc/virtual/2025/poster/30127 abstract： In the field of image editing, three core challenges persist: controllability, background preservation, and efficiency. Inversion-based methods rely on time-consuming optimization to preserve the features of the initial images, which results in low efficiency due to the requirement for extensive network inference. Conversely, inversion-free methods lack theoretical support for background similarity, as they circumvent the issue of maintaining initial features to achieve efficiency. As a consequence, none of these methods can achieve both high efficiency and background consistency. To tackle the challenges and the aforementioned disadvantages, we introduce PostEdit, a method that incorporates a posterior scheme to govern the diffusion sampling process. Specifically, a corresponding measurement term related to both the initial features and Langevin dynamics is introduced to optimize the estimated image generated by the given target prompt. Extensive experimental results indicate that the proposed PostEdit achieves state-of-the-art editing performance while accurately preserving unedited regions. Furthermore, the method is both inversion- and training-free, necessitating approximately 1.5 seconds and 18 GB of GPU memory to generate high-quality results.

## 2600. AniSDF: Fused-Granularity Neural Surfaces with Anisotropic Encoding for High-Fidelity 3D Reconstruction

链接：https://iclr.cc/virtual/2025/poster/27938 abstract： Neural radiance fields have recently revolutionized novel-view synthesis and achieved high-fidelity renderings. However, these methods sacrifice the geometry for the rendering quality, limiting their further applications including relighting and deformation. How to synthesize photo-realistic rendering while reconstructing accurate geometry remains an unsolved problem. In this work, we present AniSDF, a novel approach that learns fused-granularity neural surfaces with physics-based encoding for high-fidelity 3D reconstruction. Different from previous neural surfaces, our fused-granularity geometry structure balances the overall structures and fine geometric details, producing accurate geometry reconstruction. To disambiguate geometry from reflective appearance, we introduce blended radiance fields to model diffuse and specularity following the anisotropic spherical Gaussian encoding, a physics-based rendering pipeline. With these designs, AniSDF can reconstruct objects with complex structures and produce high-quality renderings. Furthermore, our method is a unified model that does not require complex hyperparameter tuning for specific objects. Extensive experiments demonstrate that our method boosts the quality of SDF-based methods by a great scale in both geometry reconstruction and novel-view synthesis.