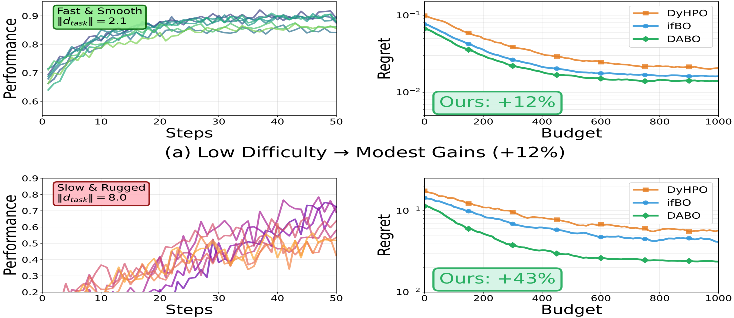
DABO: Difficulty-Aware Bayesian Optimization with Diffusion-Learned Priors

CVPR 2026 Highlight

Mengyang Li

Tianjin Normal University

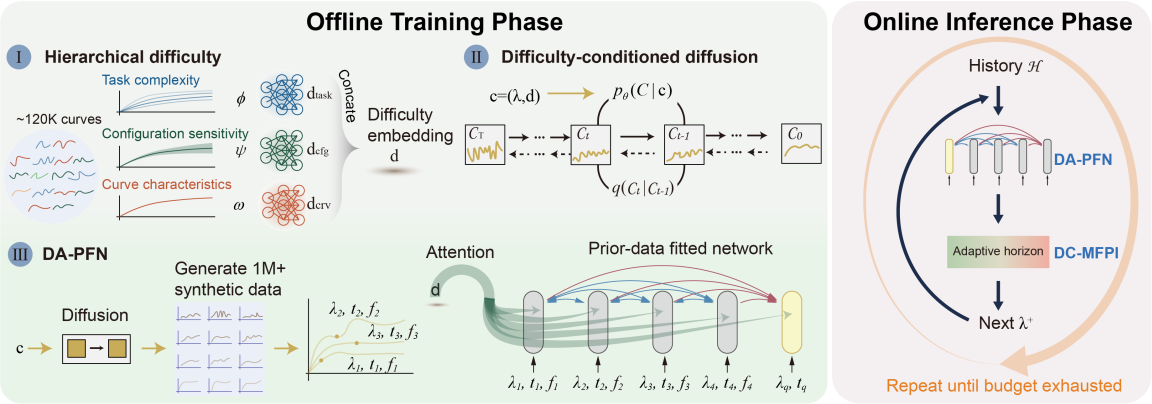
Key observation: Difficulty-agnostic methods work reasonably well on smooth landscapes, but fail catastrophically on rugged, high-difficulty search spaces.



(a) Low Difficulty → Modest Gains (+12%)

(b) High Difficulty → Large Gains (+43%)

*DABO explicitly models search-space complexity, achieving **43% lower regret** on hard tasks.*



Offline: Extract hierarchical difficulty descriptors; train a conditional diffusion model on 120K real learning curves; train the difficulty-aware surrogate DA-PFN.

Online: Guide Freeze-Thaw exploration with the difficulty-conditioned acquisition function DC-MFPI.

Hierarchical Difficulty Modeling (1/2)

We characterize search-space difficulty at three levels using explicit features, with no black-box encoder:

1. **Task complexity (ϕ): macro-level search space**
   - Performance range: $\phi_1 = \max_i \mathcal{C}_i[b_i] - \min_i \mathcal{C}_i[b_i]$
   - Trajectory diversity via DTW distance
2. **Config sensitivity (ψ): meso-level local ruggedness**
   - k-NN variance in the hyperparameter space measures local landscape roughness.
   - Local variance: $\psi_1(\lambda) = \mathrm{Std}_{j \in \mathcal{N}_k(\lambda)}[\mathcal{C}_j[b_{\max}]]$

Hierarchical Difficulty Modeling (2/2)

3. Curve dynamics (ω): micro-level evolution signals

- Improvement slope: $\omega_1(\mathcal{C}, t) = \frac{\mathcal{C}[t]-\mathcal{C}[1]}{t}$
- Oscillation volatility: $\omega_2(\mathcal{C}, t) = \text{Std}/\text{Mean}$

Unified Difficulty Descriptor

The above features are mapped by an MLP into a 192-dim vector:

$$\mathbf{d}(\lambda, t) = [\mathbf{d}_{\text{task}}; \mathbf{d}_{\text{cfg}}(\lambda); \mathbf{d}_{\text{crv}}(t)] \in \mathbb{R}^{192}$$

This vector **d** conditions both the diffusion prior and the surrogate model.

Diffusion-Learned Prior: Beyond Parametric Formulas

Prior methods (e.g., ifBO) rely on hand-crafted formulas (e.g., $y = a - bx^{-c}$) to generate synthetic curves. We replace these with a **data-driven diffusion model**.

- **Training data:** 120K real learning curves from HPO-Bench and related benchmarks.
- **Conditioning:** The difficulty descriptor $\mathbf{d}$ is injected via cross-attention.

$$\mathcal{L}_{\text{diff}} = \|\epsilon - \epsilon_\theta(\mathcal{C}_t, t, \lambda, \mathbf{d})\|^2$$

Result: 1M synthetic curves with explicit difficulty labels; ***2.3× better fidelity*** *(FCD) vs. parametric priors.*

Difficulty-Aware Surrogate: DA-PFN

Trained on 1M high-fidelity curves, DA-PFN performs fast amortized Bayesian inference.

Difficulty-Modulated Attention: We modify the standard self-attention with a multiplicative difficulty-distance factor:

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{D}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} \odot S\right)\mathbf{V}$$

where $S_{ij} = 1 + \gamma \cdot \exp(-\|\mathbf{D}_i - \mathbf{D}_j\|_2)$.

Effect: the model preferentially attends to historical observations that share the same difficulty level as the current query.

Online Deployment: Difficulty-Conditioned Acquisition

During search, we allocate compute via two difficulty-adaptive mechanisms:

1. **Adaptive horizon $h(\mathbf{d})$:** decays exponentially with config difficulty.

$$h(\mathbf{d}) = \max\left(1, \left\lfloor b_{\max} \cdot \exp\left(-\alpha \cdot \|\mathbf{d}_{\text{cfg}}\|_2\right)\right\rfloor\right)$$

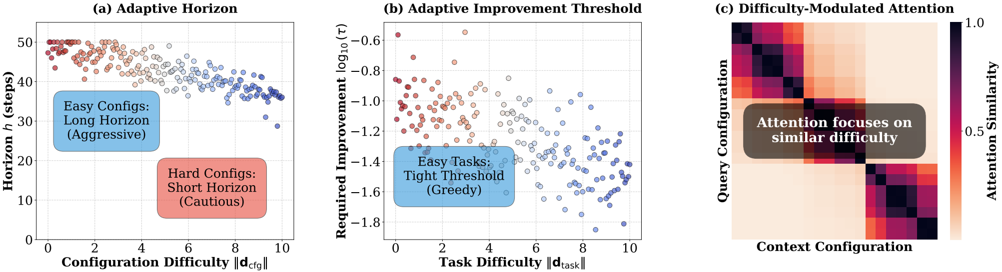
Hard regions: short horizon (cautious). Easy regions: long horizon (aggressive).

2. **Adaptive threshold $T(\mathbf{d})$:** widens with task complexity.

$$T(\mathbf{d}) = f_{\text{best}} + \tau(\|\mathbf{d}_{\text{task}}\|) \cdot (1 - f_{\text{best}})$$

Complex tasks: wider tolerance encourages exploration to escape local optima.

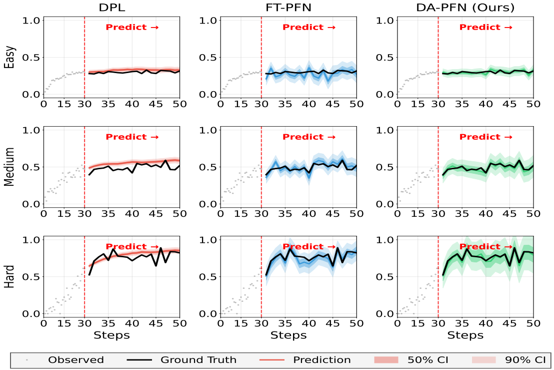
Visualization: Adaptive Mechanisms in Action



- **(a) Prediction horizon:** hard configs → very short horizon (highly cautious).
- **(b) Tolerance threshold:** complex tasks → wider threshold (encourages exploration).
- **(c) Attention weights:** strong attention appears along the difficulty-matching diagonal (highlighted in red).

Surrogate Prediction Quality: Calibration

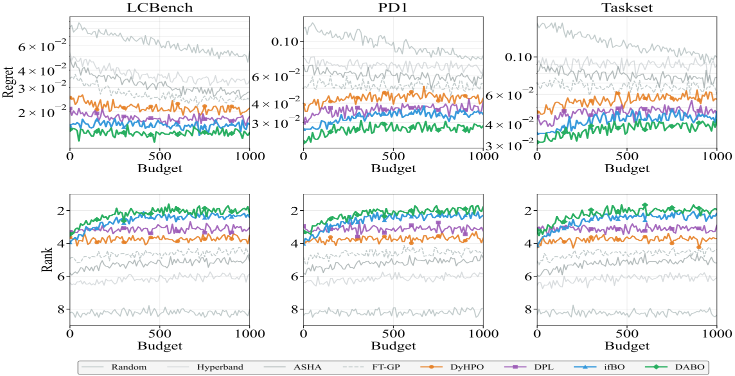
How well does DA-PFN predict? A figure says more than words.



- **DPL:** severely overconfident; predictions collapse on hard curves.
- **ifBO (FT-PFN):** fails to capture volatility on difficult curves.
- **DA-PFN (Ours):** uncertainty bands (green shading) tightly and correctly wrap the true trajectory.

End-to-End HPO Performance

Evaluated on **75 deep learning tasks** across LCBench, PD1, and Taskset.



- **11–18% lower regret** vs. prior SOTA (ifBO) on average.
- **>20% improvement** on high-difficulty tasks (e.g., ImageNet-ResNet, 8-param Transformer).

Ablation Study: Where Do the Gains Come From?

Ablation on LCBench. Each component contributes independently:

Diffusion Prior	**DA-PFN**	**Adaptive Acquisition**	**Regret** $\downarrow$
			0.0160 ± 0.002 (ifBO)
✓			0.0152 ± 0.002
	✓		0.0148 ± 0.002
		✓	0.0154 ± 0.002
✓	✓	✓	$\mathbf{0.0140 \pm 0.001}$

- **Diffusion prior alone:** 5.0% reduction (data-driven curves beat hand-crafted formulas).
- **DA-PFN alone:** 7.5% reduction (difficulty-targeted inference is highly effective).
- **All three combined:** synergistic gains, **12.5% total reduction**.

Summary

From Difficulty-Agnostic to Difficulty-Aware HPO

- **Key insight:** Search spaces have heterogeneous difficulty; treating them uniformly wastes compute.
- **Hierarchical difficulty descriptor:** Captures macro/meso/micro difficulty without black-box encoders.
- **Diffusion prior:** Data-driven curve generation far outperforms parametric alternatives ($2.3\times$ better FCD).
- **DA-PFN + adaptive acquisition:** Difficulty-modulated attention and adaptive horizon/threshold deliver consistent gains.
- **Result:** 11–18% lower regret on average; up to **43% on hard tasks**.

Thank You

Questions & Discussion

limengyang@tjnu.edu.cn