Supplementary Material

1 Computational Overhead

Module/Eq.	Complexity/$O(\cdot)$	Total FLOPs	Peak RAM
Graph Construction (1)	$N_c^2 \cdot D$	7.41×10^{14}	**778** MB
Spectral Embedding (2)	N_c^3	2.10×10^{12}	163 KB
Cost Matrix (3)	$M_c \cdot N_c \cdot d$	6.30×10^{12}	655 KB
Hungarian Match	$M_c \cdot N_c^2$	8.40×10^{13}	741 KB
PD-CFD Loss (9)	$(N_c + M_c)N_{freq}d$	9.20×10^{12}	1.03 MB
Total FAST	–	8.43×10^{14}	$\approx$ 782 MB
ResNet-50 Comparison	–	5.25×10^{15}	$\approx$ 6.25 GB

Table 1: Expensive Computation Breakdown on ImageNet-1K.

Note: FAST assumes per-class selection. ResNet-50 FLOPs are for one epoch inference on ImageNet-1K with a batch size of 256. Total feature extraction overhead of FAST on ImageNet-1K is 16% of ResNet-50.

Keep Rate	**10%**	**Time(h)**	**20%**	**Time(h)**	**30%**	**Time(h)**	**Full**
previous SOTA‡	53.6	7.9	62.6	21.1	64.5	30.2	71.2
FAST‡	**56.2**	**2.5**	**63.9**	**6.1**	**66.9**	**8.2**	71.2

Table 2: Performance Comparison on ImageNet-1K. ‡On CPU.

2 Optimization Stability

The *Hungarian mapping* evolves from dynamic exploration to stable anchoring, ensuring robust continuous-to-discrete alignment. Despite discrete updates to π, the loss maintains continuity at cost-equalizing *Voronoi boundaries*, and divergence is prevented by topological constraints that limit gradient fluctuations to the local manifold neighborhood. Since *Hungarian algorithm* minimizes the transport cost globally, gradient descent on the smooth objective $\tilde{Y}$ ensures $\mathscr{L}(\tilde{Y}^{(t+1)}, \pi^{(t+1)}) \leq \mathscr{L}(\tilde{Y}^{(t)}, \pi^{(t)})$, yielding monotonic descent optimization: $0 \leq \mathscr{L}_{total}^{(t+1)} \leq \mathscr{L}_{total}^{(t)}$, where *Monotone Convergence Theorem* guarantees convergence.

3 Necessity of Embeddings

In LLM experiments, we utilize general Sentence-BERT embeddings as standard preprocessing for manifold construction. The structural distinction between the encoder-only embedder and decoder-only LLM introduces a general semantic prior rather than coupled architectural bias, confirmed by Table **??** that shows FAST's robust generalization.

Model	**L-FAST**	**L-Rand**	**Q-FAST**	**Q-Rand**	**M-FAST**	**M-Rand**
Acc (%)	$39.0_{\uparrow 8.7}$	$35.9_{\uparrow 5.6}$	$72.3_{\uparrow 6.1}$	$70.9_{\uparrow 4.7}$	$63.1_{\uparrow 4.0}$	$61.6_{\uparrow 2.5}$

Table 3: Supplementary Experiments on LLM datasets.

Notation: L: LLaMA2-7B, Q: Qwen2.5-7B, M: Mistral-7B, Rand:Random. $\uparrow$ indicates the relative improvement to the base model.

4 More Ablation Studies

Use CIFAR-10 dataset (10% KR, training ResNet-18) unless noted; random sampling shows $75.70\%_{\pm 7.63}$ accuracy.

Method	Acc (%)	Time (s)	Energy (Wh)	Std	Opt/Steps
Pixel-Opt‡	61.52	3770	29.75	13.73	1300
ResNet-50†	81.51	627.67	37.75	3.32	230
FAST‡	**90.32**	**353**	**1.409**	**1.21**	**80**

Table 4: Graph Feature Extractor Ablation on ‡CPU, †GPU.

Analysis: Pixel-level optimization fails due to the sparsity and noise of high-dimensional space, while incurring massive computational costs (10× slower). ResNet-50 features suffer from architectural bias and computational overhead.

Parameter	Ours	$k=5$	$k=50$	$d=8$	$d=128$
Trustworthiness	$\mathbf{0.95}_{\pm \mathbf{0.01}}$	$0.73_{\pm 0.10}$	$0.52_{\pm 0.09}$	$0.44_{\pm 0.09}$	$0.92_{\pm 0.02}$
Continuity	$0.91_{\pm 0.03}$	$0.49_{\pm 0.15}$	$0.93_{\pm 0.06}$	$0.92_{\pm 0.03}$	$0.87_{\pm 0.06}$
Acc (%)	$\mathbf{90.32}_{\pm \mathbf{1.21}}$	$72.12_{\pm 5.56}$	$85.01_{\pm 3.27}$	$78.86_{\pm 3.13}$	$86.73_{\pm 4.16}{}^{*}$

Table 5: Hyperparameter Ablation (neighbor scale k and reduced dimension d). *Time doubles. FAST adopts $k=15$ and $d=32$.

Analysis: Small k fractures the manifold, while large k causes collapse, connecting distant classes; low d bottlenecks information, and high d adds noise and doubles optimization time. Our settings optimally preserve topology.

Strategy	DAS	PES	PUS	US	Collinear	PDAS
Acc (%)	$88.89_{\pm 2.23}$	$87.12_{\pm 2.60}$	$88.66_{\pm 2.01}$	$86.60_{\pm 2.73}$	$74.16_{\pm 7.77}$	$\mathbf{90.32}_{\pm \mathbf{1.21}}$
Opt/Steps	150	400	250	430	100^{*}	**80**

Table 6: Impact of Frequency Selection. *Optimization failure mode. Refer to paper Fig. 9 for illustrations.

Analysis: Baselines unstably converge due to suboptimal frequency selection; PDAS achieves optimal stability by progressively selecting discriminative frequencies.

Method	CIFAR-10	RESISC-45	DTD
FAST-Vanilla	$88.17_{\pm 2.12}$	$82.01_{\pm 3.99}$	$41.15_{\pm 3.20}$
FAST-PD	$\mathbf{90.32}_{\pm \mathbf{1.21}}$	$\mathbf{85.00}_{\pm \mathbf{2.01}}$	$\mathbf{45.77}_{\pm \mathbf{2.53}}$

Table 7: Phase-Decoupled CFD Ablation (full ablation in Fig. 7).

Analysis: *CFD* fails to capture high-frequency details due to amplitude-phase coupling. *PD-CFD* resolves this, showing substantial gains on texture-rich DTD (+4.62%) and RESISC-45 (+2.99%) compared to CIFAR-10 (+2.15%).

5 Additional Analysis on Limitations of KL and CE for High-Order Moment Alignment

Notation. Let P be an unknown target distribution on $\mathscr{X} \subseteq \mathbb{R}^d$ with density p. An exponential family is

$$\mathscr{Q}_T := \left\{ q_\theta(x) = h(x) \exp\big(\theta^\top T(x) - A(\theta)\big) \;\middle|\; \theta \in \Theta \subset \mathbb{R}^m \right\},$$

where $T : \mathscr{X} \to \mathbb{R}^m$ is a vector of sufficient statistics, $A(\theta) := \log \int h(x) \exp(\theta^\top T(x))\, dx$ is the log-partition function (cumulant generating function), and h is a base density. We write $\mu(\theta) := \nabla A(\theta) = \mathbb{E}_{q_\theta}[T(X)]$ for the mean-parameter map.

For $k \in \mathbb{N}$, define the (truncated) polynomial statistics

$$T^{(\le k)}(x) \;=\; \big(x,\, xx^\top,\, x^{\otimes 3},\, \ldots,\, x^{\otimes k}\big),$$

so that $\mathscr{Q}_{\le k}$ denotes the exponential family whose sufficient statistics are all monomials up to degree k (with a suitable choice of h to ensure normalizability).

Proposition 1 (I-projection yields moment matching on the chosen statistics)**.** *Consider the KL minimization (I-projection) of P onto an exponential family $\mathscr{Q}_T$:*

$$\theta^\star \;\in\; \arg\min_{\theta\in\Theta} D_{\mathrm{KL}}\big(P \,\|\, Q_\theta\big) \;=\; \arg\min_{\theta} \int p(x) \log \frac{p(x)}{q_\theta(x)}\, dx. \tag{1}$$

Then A is convex and the KL objective is strictly convex in θ on the mean-parameter interior; the unique minimizer $\theta^\star$ (when it exists) satisfies the first-order optimality condition

$$\nabla_\theta\Big(\mathbb{E}_P[\theta^\top T(X)] - A(\theta)\Big)\Big|_{\theta=\theta^\star} \;=\; \mathbb{E}_P[T(X)] - \nabla A(\theta^\star) \;=\; 0.$$

Equivalently,

$$\mathbb{E}_P[T(X)] \;=\; \mathbb{E}_{Q_{\theta^\star}}[T(X)]\,.$$

Proof (via convex duality and Fenchel–Legendre conjugacy). Expanding the KL divergence,

$$D_{\mathrm{KL}}(P\|Q_\theta) = -H(P) - \mathbb{E}_P\big[\theta^\top T(X) - A(\theta) + \log h(X)\big],$$

where $H(P)$ and $\mathbb{E}_P[\log h(X)]$ are θ-independent constants. Thus minimizing $D_{\mathrm{KL}}(P\|Q_\theta)$ is equivalent to maximizing the concave functional

$$\mathscr{L}(\theta) := \mathbb{E}_P[\theta^\top T(X)] - A(\theta).$$

Since A is convex (indeed, A is the log-moment generating function of T under h), $\mathscr{L}$ is concave and (under standard interiority conditions) has a unique maximizer $\theta^\star$. The Fenchel–Legendre conjugate of A is $A^*(\mu) := \sup_\theta\{\langle\theta,\mu\rangle - A(\theta)\}$; strong duality yields that the optimal $\theta^\star$ satisfies $\nabla A(\theta^\star) = \mu^\star = \mathbb{E}_P[T(X)]$ (i.e. the moment-matching equations). Finally $\nabla A(\theta) = \mathbb{E}_{Q_\theta}[T(X)]$ by standard exponential-family calculus, completing the proof. □

Corollary 1 (Distributions controlled by KL under restricted sufficient statistics)**.** *Let $\mathscr{Q}_T$ be any exponential family with sufficient statistics T. Then the I-projection $Q_{\theta^\star}$ aligns exactly those coordinates of T:*

$$\mathbb{E}_P[T_i(X)] = \mathbb{E}_{Q_{\theta^\star}}[T_i(X)] \quad \textit{for each component } T_i,$$

but imposes no necessary constraint on expectations of functions f lying outside the linear span of $\{1, T_1, \ldots, T_m\}$. In particular, if $T = T^{(\le k)}$ contains all monomials up to degree k, then KL minimization guarantees matching of all raw moments up to order k, and does not *in general constrain any $(k+1)$-st or higher-order moments.*

Proof. Immediate from Proposition 1, noting that the optimality system is exactly the linear system matching $\mathbb{E}_P[T]$ and $\mathbb{E}_{Q_\theta}[T]$. Any f outside the closed linear span of $\{1, T\}$ cannot be represented as $\alpha_0 + \alpha^\top T$, hence its expectation is unconstrained by the KKT system. □

Corollary 2 (Gaussian family ($k=2$) controls mean and covariance but not higher cumulants)**.** *Let $\mathscr{Q}_{\mathscr{N}} = \{\mathscr{N}(\mu,\Sigma)\}$. Then $\mathscr{Q}_{\mathscr{N}}$ is an exponential family with $T(x) = (x, xx^\top)$ and $\nabla A(\mu,\Sigma) = (\mathbb{E}[X], \mathbb{E}[XX^\top])$. The I-projection $Q_{\theta^\star} = \mathscr{N}(\mu^\star,\Sigma^\star)$ thus satisfies*

$$\mathbb{E}_{Q_{\theta^\star}}[X] = \mathbb{E}_P[X], \qquad \mathbb{E}_{Q_{\theta^\star}}[XX^\top] = \mathbb{E}_P[XX^\top],$$

but neither skewness (third cumulant) nor kurtosis (fourth cumulant) is constrained or minimized by the KL objective in general.

Proof. A direct specialization of Corollary 1 with $k = 2$. □

Proposition 2 (Augmenting statistics raises the matched moment order, but only up to that order)**.** *Let $\mathscr{Q}_{\le k}$ denote the exponential family with $T^{(\le k)}$. Then the sequence of I-projections $\{Q^{(k)}\}_{k\ge 1}$ defined by*

$$Q^{(k)} \in \arg\min_{Q\in\mathscr{Q}_{\le k}} D_{\mathrm{KL}}(P\|Q)$$

satisfies, for each fixed k,

$$\mathbb{E}_{Q^{(k)}}[X^{\otimes r}] \;=\; \mathbb{E}_P[X^{\otimes r}] \quad \textit{for all } r = 1,2,\ldots,k,$$

and there is no general guarantee that $\mathbb{E}_{Q^{(k)}}[X^{\otimes r}]$ aligns with $\mathbb{E}_P[X^{\otimes r}]$ for any $r > k$.

Proof. By Proposition 1 applied to $T^{(\le k)}$, the optimality conditions enforce equality of the first k raw-moment tensors. Since the KL objective reduces to a linear functional of T minus $A(\theta)$, any statistics not included in T do not appear in the optimality system and remain uncontrolled. □

Remark 1 (Consequences for KL/CE-based *evaluation* of distributional alignment)**.** *Since cross-entropy minimization is equivalent to minimizing $D_{\mathrm{KL}}(P\|Q_\theta)$ when P is fixed, Propositions 1–2 imply a fundamental limitation:* Under a restricted exponential family, KL/CE can certify alignment of at most the moments encoded in T (e.g., mean and covariance for Gaussians), and it provides no guarantees for higher-order moments or cumulants that are not present in T. *Hence, unless higher-order statistics are explicitly included in the model (i.e., increasing k), KL/CE-based fitting and evaluation do not ensure moment alignment.*

Example 1 (Gaussian vs. heavy-tailed target: kurtosis mismatch at the KL optimum)**.** *Let P be a zero-mean heavy-tailed distribution (e.g., a centered Laplace in 1D) with variance σ^2 and kurtosis $\kappa_P > 3$. Its I-projection onto $\mathscr{Q}_{\mathscr{N}}$ is $Q^\star = \mathscr{N}(0,\sigma^2)$, which matches mean and variance by Corollary 2. However, $\kappa_{Q^\star} = 3 \neq \kappa_P$, i.e., fourth-order moments are not aligned at the KL optimum—a concrete manifestation of Remark 1.*

Remark 2 (Mixtures vs. single exponential families)**.** *A Gaussian mixture model (GMM) is a convex combination of Gaussians and* not *a single exponential family with a fixed finite-dimensional T. Therefore, the structural limitation in Propositions 1–2 does not apply in the same form to GMMs: with sufficiently many components, a GMM can approximate higher-order structures arbitrarily well. However, practical optimization remains nonconvex and capacity-limited, so alignment may still fail in practice despite the absence of a finite-dimensional T.*

A General Counterexample: Small KL yet Divergent Higher-Order Moments

We now give a short, self-contained construction showing that (beyond model restrictions) *the KL divergence itself does not control unbounded test-function expectations*, such as moments.

Proposition 3 (KL can be arbitrarily small while k-th moments diverge)**.** *Fix any integer $k \ge 3$. There exists a sequence of distributions P_M and a reference Q such that*

$$D_{\mathrm{KL}}(P_M\|Q) \to 0 \quad \textit{but} \quad \big|\mathbb{E}_{P_M}[X^k] - \mathbb{E}_Q[X^k]\big| \to \infty.$$

Proof. Let $Q = \mathscr{N}(0,1)$ and $R_M = \mathscr{N}(M,1)$. Define the mixture

$$P_M = (1-\varepsilon_M)\,Q + \varepsilon_M\,R_M, \qquad \varepsilon_M = \frac{1}{M^2\log M}.$$

(i) KL upper bound by joint convexity. By convexity of $D_{\mathrm{KL}}(\cdot\|Q)$ in its first argument,

$$D_{\mathrm{KL}}\big(P_M\|Q\big) \le (1-\varepsilon_M)\cdot D_{\mathrm{KL}}(Q\|Q) + \varepsilon_M \cdot D_{\mathrm{KL}}(R_M\|Q) = \varepsilon_M \cdot \tfrac{M^2}{2} = \tfrac{1}{2\log M} \to 0, \tag{2}$$

where $D_{\mathrm{KL}}(\mathscr{N}(M,1)\|\mathscr{N}(0,1)) = \frac{M^2}{2}$.

(ii) Divergence of the k-th moment gap. For $k \ge 3$, $\mathbb{E}_{R_M}[X^k] \sim M^k$. Hence

$$\big|\mathbb{E}_{P_M}[X^k] - \mathbb{E}_Q[X^k]\big| = \varepsilon_M\big|\mathbb{E}_{R_M}[X^k] - \mathbb{E}_Q[X^k]\big| \gtrsim \varepsilon_M M^k = \frac{M^{k-2}}{\log M} \xrightarrow[M\to\infty]{} \infty. \tag{3}$$

Thus, KL can be made arbitrarily small while the k-th moment discrepancy diverges. □

Interpretation. Pinsker's inequality controls only *bounded* test functions via total variation; polynomial moments are unbounded, so small KL does not imply closeness of higher moments. Proposition 3 complements the structural limitation (Propositions 1–2) by showing a *metric* limitation: even without any modeling restrictions, KL does not bound higher-order moments.

6 Limitations of MMD for Aligning Higher-Order Statistics

Maximum Mean Discrepancy (MMD) is widely used as a nonparametric distributional metric, defined for a positive definite kernel k by

$$\mathrm{MMD}_k(P,Q) = \|\mu_P - \mu_Q\|_{\mathscr{H}_k} = \sup_{\|f\|_{\mathscr{H}_k}\le 1} \left(\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]\right), \tag{4}$$

where $\mathscr{H}_k$ is the reproducing kernel Hilbert space (RKHS) associated with k. When k is characteristic, MMD metrizes weak convergence. However, for the purpose of coreset selection—where the objective is to align *all* statistical moments (including heavy-tailed or high-order cumulants)— weak convergence is insufficient. We show below that with the commonly used bounded kernels (e.g., Gaussian, Laplacian), MMD does *not* control higher-order moments, even when $\mathrm{MMD}_k(P,Q)$ is arbitrarily small.

Setup. Let $k(x,y)$ be any bounded positive definite kernel with $\sup_{x,y}|k(x,y)| \le K < \infty$ (this includes the Gaussian RBF). Let Q be any reference distribution with finite moments of all orders.

We construct the following sequence:

$$P_M = (1-\varepsilon_M)Q + \varepsilon_M R_M,$$

where R_M is a distribution concentrated at radius $\|x\| \approx M$, and ε_M is a vanishing mixing weight.

The construction is analogous to the KL example in Proposition 3 but adapted to the MMD geometry.

Proposition 4 (Small MMD does not control higher-order moments)**.** *For any bounded kernel k and any integer $r \ge 3$, there exists a sequence P_M such that*

$$\mathrm{MMD}_k(P_M,Q) \longrightarrow 0 \quad \textit{but} \quad \left|\mathbb{E}_{P_M}[X^r] - \mathbb{E}_Q[X^r]\right| \longrightarrow \infty.$$

Proof. Because k is bounded, the RKHS norm of the kernel mean embedding satisfies

$$\|\mu_{P_M} - \mu_Q\|_{\mathscr{H}_k} \le \mathbb{E}_{P_M,Q}[\,|k(X,Y)|\,] \le K.$$

More precisely,

$$\mu_{P_M} = (1-\varepsilon_M)\mu_Q + \varepsilon_M \mu_{R_M},$$

hence

$$\mathrm{MMD}_k(P_M,Q) = \varepsilon_M \|\mu_{R_M} - \mu_Q\|_{\mathscr{H}_k} \le 2K\,\varepsilon_M.$$

Taking $\varepsilon_M = 1/\log M$ yields

$$\mathrm{MMD}_k(P_M,Q) \le \frac{2K}{\log M} \xrightarrow[M\to\infty]{} 0.$$

Next, choose R_M such that almost all its mass lies on $\|x\| \approx M$. Then the r-th moment satisfies

$$\mathbb{E}_{R_M}[X^r] \asymp M^r,$$

and therefore

$$\mathbb{E}_{P_M}[X^r] = (1-\varepsilon_M)\mathbb{E}_Q[X^r] + \varepsilon_M\,\mathbb{E}_{R_M}[X^r] \asymp \mathbb{E}_Q[X^r] + \varepsilon_M M^r. \tag{5}$$

Since $\varepsilon_M M^r = M^r/\log M \to \infty$ for any $r \ge 3$, we obtain

$$\left|\mathbb{E}_{P_M}[X^r] - \mathbb{E}_Q[X^r]\right| \longrightarrow \infty,$$

even while $\mathrm{MMD}_k(P_M,Q) \to 0$. □

Interpretation. This result parallels Proposition 3 for KL: both reveal a *metric limitation* independent of modeling assumptions. For bounded kernels, MMD metrizes weak convergence but cannot control expectations of unbounded test functions—particularly polynomial functions that capture higher-order moments, tail behavior, or heavy-tailed anisotropy.

Consequently, when MMD is used as an objective for coreset construction, moment alignment depends critically on the expressiveness of the chosen kernel. If the kernel is insufficient to probe heavy tails or high-order interactions, the resulting coreset may perfectly match the MMD score yet deviate drastically in higher-order statistics—precisely the regime where our PD-CFD metric provides a more faithful discrepancy measure.

7 Characteristic Functions: Complete Statistical Representation and Fourier–Analytic Foundations

Notation. Let P be a Borel probability measure on $\mathbb{R}^d$ with random vector $X \sim P$. The *characteristic function* (CF) of P is

$$\varphi_P(\omega) \;:=\; \int_{\mathbb{R}^d} e^{i\omega^\top x} P(dx) \;=\; \mathbb{E}_P[e^{i\omega^\top X}], \qquad \omega \in \mathbb{R}^d.$$

Each φ_P is bounded ($|\varphi_P| \le 1$), uniformly continuous, and positive definite.

Proposition 5 (Fourier inversion: recovering P from φ_P)**.** *Suppose P admits a density $p \in L^1(\mathbb{R}^d)$. Then p can be recovered from its characteristic function by the inverse Fourier transform:*

$$p(x) \;=\; \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\omega^\top x}\, \varphi_P(\omega)\, d\omega, \qquad \text{for a.e. } x \in \mathbb{R}^d.$$

More generally, for any $f \in L^1(\mathbb{R}^d)$ with Fourier transform $\widehat{f}(\omega) := \int e^{i\omega^\top x} f(x)\, dx$,

$$\int_{\mathbb{R}^d} f(x)\, P(dx) \;=\; \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \widehat{f}(\omega)\, \varphi_P(\omega)\, d\omega.$$

Thus, the knowledge of φ_P on $\mathbb{R}^d$ uniquely determines all integrals $\int f\, dP$, and hence the measure P itself.

Proof (Fourier transform duality). The Fourier transform $\mathscr{F} : f \mapsto \widehat{f}$ is a linear bijection between L^1 and bounded continuous functions, satisfying the Plancherel identity $\int |f|^2 = (2\pi)^{-d} \int |\widehat{f}|^2$ for $f \in L^2$. Since $e^{i\omega^\top x}$ is the kernel of this transform,

$$\int f(x)\, P(dx) \;=\; \int f(x) \left(\frac{1}{(2\pi)^d} \int e^{-i\omega^\top x}\, \varphi_P(\omega)\, d\omega \right) dx \;=\; \frac{1}{(2\pi)^d} \int \widehat{f}(\omega)\, \varphi_P(\omega)\, d\omega. \tag{6}$$

If φ_P is known everywhere, the right-hand side gives $\int f\, dP$ for all $f \in L^1 \cap L^2$, implying uniqueness of P. □

Proposition 6 (Lévy's continuity theorem: weak convergence and uniqueness)**.** *Let $\{P_n\}$ be a sequence of probability measures on $\mathbb{R}^d$ with characteristic functions φ_{P_n}. Suppose that $\varphi_{P_n}(\omega) \to \varphi(\omega)$ for each ω, and that the limit φ is itself a characteristic function (i.e. positive definite and $\varphi(0) = 1$). Then $P_n \Rightarrow P$, where P is the distribution with CF φ. In particular, if $\varphi_P(\omega) \equiv \varphi_Q(\omega)$ for all ω, then $P = Q$.*

Proof (weak convergence via Fourier test functions). For any $f \in C_c^\infty(\mathbb{R}^d)$, its Fourier transform $\widehat{f}$ is rapidly decaying. By Fubini and dominated convergence,

$$\int f(x)\, P_n(dx) \;=\; \frac{1}{(2\pi)^d} \int \widehat{f}(\omega)\, \varphi_{P_n}(\omega)\, d\omega \;\xrightarrow[n\to\infty]{}\; \frac{1}{(2\pi)^d} \int \widehat{f}(\omega)\, \varphi(\omega)\, d\omega \;=\; \int f(x)\, P(dx). \tag{7}$$

Hence $\int f\, dP_n \to \int f\, dP$ for all $f \in C_c^\infty$, which implies $P_n \Rightarrow P$ by the Portmanteau theorem. Taking $P_n = P$ and $\varphi_P = \varphi_Q$ gives the uniqueness $P = Q$. □

Proposition 7 (Bochner's theorem: positive-definite functions as Fourier transforms of measures)**.** *A continuous function $\psi : \mathbb{R}^d \to \mathbb{C}$ with $\psi(0) = 1$ is the characteristic function of some probability measure on $\mathbb{R}^d$ if and only if it is* positive definite*, i.e. for all $n \in \mathbb{N}$, all $\omega_1, \ldots, \omega_n \in \mathbb{R}^d$, and all $c_1, \ldots, c_n \in \mathbb{C}$,*

$$\sum_{i,j=1}^{n} c_i \overline{c_j}\, \psi(\omega_i - \omega_j) \geq 0.$$

Moreover, for every finite nonnegative measure μ on $\mathbb{R}^d$, its Fourier transform $\psi(\omega) := \int e^{i\omega^\top x} d\mu(x)$ is continuous and positive definite. Thus, continuous positive-definite functions with $\psi(0) = 1$ are in one-to-one correspondence with characteristic functions of probability measures.

Proof (spectral representation). For any finite positive measure μ, $\psi(\omega) = \int e^{i\omega^\top x} d\mu(x)$ is continuous and satisfies the positive-definite inequality above by direct calculation. Conversely, if ψ is continuous and positive definite, by the classical Bochner–Khintchine theorem there exists a unique finite nonnegative measure μ such that ψ is its Fourier transform. Setting $\mu(\mathbb{R}^d) = 1$ gives a probability measure $P = \mu$, with $\psi = \varphi_P$. □

Proposition 8 (Matching characteristic functions on $\mathbb{R}^d$ equals to matching the joint distribution)**.** *Let P, Q be Borel probability measures on $\mathbb{R}^d$ with characteristic functions φ_P, φ_Q. Define a nonnegative weighting function $w \in L^1(\mathbb{R}^d)$ satisfying* $\mathrm{supp}(w) = \mathbb{R}^d$ *and the frequency-domain distance*

$$\mathscr{D}_w^2(P, Q) \;:=\; \int_{\mathbb{R}^d} w(\omega) \left|\varphi_P(\omega) - \varphi_Q(\omega)\right|^2 d\omega.$$

Then the following are equivalent:

$$\mathscr{D}_w(P, Q) = 0 \quad \Longleftrightarrow \quad \varphi_P(\omega) = \varphi_Q(\omega)\ \forall \omega \quad \Longleftrightarrow \quad P = Q. \tag{8}$$

Proof (Lévy–Bochner synthesis). If $\mathscr{D}_w(P, Q) = 0$, then $\varphi_P = \varphi_Q$ almost everywhere on $\{w > 0\}$. Since both φ_P, φ_Q are continuous, equality extends to all $\omega \in \mathbb{R}^d$. By Proposition 6, $\varphi_P \equiv \varphi_Q$ implies $P = Q$. Conversely, if $P = Q$, then clearly $\mathscr{D}_w(P, Q) = 0$. Thus, equality of CFs on the full frequency domain is equivalent to equality of the underlying distributions. □

Intuitively, $D_w(P, Q)$ measures the squared discrepancy between characteristic functions across all frequencies, weighted by w. When w has full support, this distance captures the complete distributional difference. We now connect this frequency-domain view to kernel methods via Bochner's theorem.

Proposition 9 (Kernel formulation via Bochner's theorem)**.** *Let $k(x, y) = \kappa(x - y)$ be a bounded, continuous, translation-invariant kernel with spectral measure Λ satisfying*

$$\kappa(t) \;=\; \int_{\mathbb{R}^d} e^{i\omega^\top t}\, d\Lambda(\omega).$$

Then the maximum mean discrepancy (MMD) between P, Q in the RKHS of k admits the spectral representation

$$\mathrm{MMD}_k^2(P, Q) \;=\; \int_{\mathbb{R}^d} \left|\varphi_P(\omega) - \varphi_Q(\omega)\right|^2 d\Lambda(\omega).$$

Moreover, if the kernel is characteristic*—equivalently, the support of Λ is all of $\mathbb{R}^d$—then $\mathrm{MMD}_k(P, Q) = 0$ if and only if $P = Q$.*

Proof (spectral integration). Expanding expectations and applying Bochner's representation,

$$\begin{aligned} \mathbb{E}\, k(X, Y) \;&=\; \int_{\mathbb{R}^d} \mathbb{E}\, e^{i\omega^\top (X - Y)}\, d\Lambda(\omega) \\ &=\; \int_{\mathbb{R}^d} \varphi_P(\omega)\, \overline{\varphi_Q(\omega)}\, d\Lambda(\omega). \end{aligned} \tag{9}$$

Substituting into the definition $\mathrm{MMD}_k^2(P, Q) = \mathbb{E}\, k(X, X') + \mathbb{E}\, k(Y, Y') - 2\mathbb{E}\, k(X, Y)$ yields the stated formula. If $\mathrm{MMD}_k(P, Q) = 0$ while Λ has full support, the integrand's continuity forces $\varphi_P \equiv \varphi_Q$, which by Proposition 6 implies $P = Q$. □

Summary and Connection to CFD. Propositions 2–6 collectively establish that a probability distribution is uniquely and completely determined by its characteristic function, and that matching characteristic functions over the entire frequency domain—either directly via a weighted L^2 distance D_w or indirectly via a characteristic kernel MMD—is equivalent to matching the full joint distribution. The Fourier inversion theorem connects CFs to densities, Bochner's theorem connects positive-definite kernels to spectral measures, and Lévy's continuity theorem ensures that equality of CFs implies equality of distributions. Consequently, frequency-domain metrics such as our CFD provide a principled way to capture all moments and dependencies, beyond what marginal or covariance-based criteria can express.

8 Moments and Cumulants via Taylor Expansion of Characteristic Functions

8.1 Preliminaries and Notations

Let $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ be a random vector with law P. Its *characteristic function* (CF) is

$$\varphi(\omega) \;=\; \mathbb{E}\Big[e^{i\,\omega^\top X}\Big], \qquad \omega = (\omega_1, \ldots, \omega_d) \in \mathbb{R}^d,$$

and the *log-characteristic function* (log-CF) is $\psi(\omega) = \log \varphi(\omega)$ whenever $\varphi(\omega) \neq 0$.

Multi-index notation. A *multi-index* is $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $|\alpha| := \sum_{j=1}^d \alpha_j$, $\alpha! := \prod_{j=1}^d \alpha_j!$, $\omega^\alpha := \prod_{j=1}^d \omega_j^{\alpha_j}$, and $X^\alpha := \prod_{j=1}^d X_j^{\alpha_j}$. The mixed partial derivative is

$$\partial^\alpha \;:=\; \frac{\partial^{|\alpha|}}{\partial \omega_1^{\alpha_1} \cdots \partial \omega_d^{\alpha_d}}.$$

Standing condition (finite moment of order $|\alpha|$). Throughout, whenever we speak of $\partial^\alpha \varphi(0)$ we assume $\mathbb{E}|X^\alpha| < \infty$. This ensures, via dominated convergence, that differentiation and expectation can be interchanged.

8.2 Limitations of Marginal-Only Matching

Classical matching of *univariate* moments (per-coordinate means/variances/skewness/kurtosis) aligns only the marginals $\{P_{X_j}\}_{j=1}^d$. However, cross-variable dependence (pairwise, triple-wise, and beyond) lives in *mixed* moments such as $\mathbb{E}[X_i X_j]$, $\mathbb{E}[X_i X_j X_k]$, etc. Hence, marginal-only criteria cannot in general control the joint distribution. We next show that the CF and its derivatives at the origin provide an exact, layered access to all mixed moments/cumulants, thereby capturing dependence.

8.3 Taylor Coefficients and Mixed Moments/ Cumulants

Proposition 10 (Derivative–Moment Correspondence)**.** *Suppose $\mathbb{E}|X^\alpha| < \infty$. Then*

$$\partial^\alpha \varphi(0) \;=\; i^{|\alpha|}\, \mathbb{E}[X^\alpha]$$

for every multi-index $\alpha \in \mathbb{N}_0^d$.

Proof. Write $e^{i\,\omega^\top X} = \prod_{j=1}^d e^{i\,\omega_j X_j}$ and differentiate. Each differentiation w.r.t. ω_j contributes a factor iX_j. Thus $\partial^\alpha e^{i\,\omega^\top X} = (iX_1)^{\alpha_1} \cdots (iX_d)^{\alpha_d} e^{i\,\omega^\top X}$. Taking expectation and setting $\omega = 0$ yields the claim, with the interchange of differentiation and expectation justified by $\mathbb{E}|X^\alpha| < \infty$. □

Low-order examples. For $\alpha = e_j$ (the j-th unit vector), $\partial^\alpha \varphi(0) = i\,\mathbb{E}[X_j]$ (means). For $\alpha = e_j + e_k$, $\partial^\alpha \varphi(0) = -\,\mathbb{E}[X_j X_k]$ (mixed second moments). For $\alpha = e_i + e_j + e_k$, $\partial^\alpha \varphi(0) = -\,i\,\mathbb{E}[X_i X_j X_k]$ (triple mixed moments).

Proposition 11 (Multivariate Taylor expansion of φ at 0)**.** *Under $\mathbb{E}|X^\alpha| < \infty$ for all $|\alpha| \le m$,*

$$\varphi(\omega) \;=\; \sum_{|\alpha| \le m} \frac{1}{\alpha!}\, \partial^\alpha \varphi(0)\, \omega^\alpha \;+\; o(\|\omega\|^m) \;=\; \sum_{|\alpha| \le m} \frac{i^{|\alpha|}}{\alpha!}\, \mathbb{E}[X^\alpha]\, \omega^\alpha \;+\; o(\|\omega\|^m). \tag{10}$$

Define $\psi(\omega) = \log \varphi(\omega)$ (well-defined near 0 since $\varphi(0) = 1$ and φ is continuous).

Definition 1 (Log-Characteristic Function and Mixed Cumulants)**.** *For a multi-index $\alpha \neq 0$ with $\mathbb{E}|X^\alpha| < \infty$, the* mixed cumulant *is*

$$\kappa_\alpha := i^{-|\alpha|} \partial^\alpha \psi(0)$$

(with $\kappa_{\mathbf{0}} := 0$ by convention).

Proposition 12 (Derivative–Cumulant correspondence)**.** *Assume the moments needed are finite so that ψ is $|\alpha|$-times differentiable at 0. Then the coefficients of the multivariate Taylor series of ψ at 0 equal the mixed cumulants:*

$$\psi(\omega) = \sum_{|\alpha|\geq 1} \frac{1}{\alpha!} \partial^\alpha \psi(0)\, \omega^\alpha = \sum_{|\alpha|\geq 1} \frac{i^{|\alpha|}}{\alpha!} \kappa_\alpha\, \omega^\alpha \quad \textit{(convergent near 0).} \tag{11}$$

Proof. By the chain rule, $\nabla\psi = \nabla\varphi/\varphi$ and $\nabla^2\psi = (\nabla^2\varphi)/\varphi - (\nabla\varphi\,\nabla\varphi^\top)/\varphi^2$, etc. Evaluating at 0 and using Prop. 10:

$$\begin{aligned} &\nabla\psi(0) = i\,\mathbb{E}[X], \\ &\nabla^2\psi(0) = -\left(\mathbb{E}[XX^\top] - \mathbb{E}[X]\mathbb{E}[X]^\top\right) = -\operatorname{Cov}(X). \end{aligned} \tag{12}$$

Higher-order derivatives of ψ yield the classical cumulant tensors (via the multivariate Faà di Bruno formula). Collecting terms gives the stated Taylor expansion with coefficients $\partial^\alpha\psi(0) = i^{|\alpha|}\kappa_\alpha$. □

Interpretation and examples. $\kappa_{e_j} = \mathbb{E}[X_j]$ (means); $\kappa_{e_j+e_k} = \operatorname{Cov}(X_j, X_k)$ (covariance); third-order $\kappa_{e_i+e_j+e_k}$ measure non-Gaussian triple interactions (synergy/redundancy). In general, *cross-block independence* forces all mixed cumulants spanning the blocks to vanish, making cumulants a clean diagnostic of dependence.

Implications

- Moments: $\partial^\alpha\varphi(0) = i^{|\alpha|}\mathbb{E}[X^\alpha]$ reveals *all* mixed moments at each order $|\alpha|$.
- Cumulants: $\partial^\alpha\psi(0) = i^{|\alpha|}\kappa_\alpha$ isolates genuine interactions (they are additive for independent sums and vanish across independent groups).
- Hence, matching φ (or ψ) near 0 across *all* multi-indices matches *all* mixed moments/cumulants, aligning dependence at every order.

9 Additional Analysis on PD-CFD

Using the CUB-200-2011.[1] bird dataset illustrated in Fig. 3, we evaluate how PD-CFD improves the phase-focusing behavior of NCFM.[2]. The CUB-200-2011 dataset contains extremely fine-grained categories where discriminative information resides mainly in high-frequency components such as feather textures. Therefore, the loss must be sensitive to detailed structural variations.

We first revisit the original NCFM loss:

$$\mathrm{Chf}(t; f) = \alpha\left(\left|\Phi_{f(x)}(t) - \Phi_{\hat{f}(x)}(t)\right|^2\right) + (1-\alpha)\,\left|\Phi_{f(x)}(t)\right|\left|\Phi_{\hat{f}(x)}(t)\right|\left(1 - \cos\left(a_{f(x)}(t) - a_{\hat{f}(x)}(t)\right)\right) \tag{13}$$

where Φ denotes the characteristic function magnitude and $a(\cdot)$ the phase.

Although NCFM applies different hyperparameters to amplitude and phase, the phase term is still tightly entangled with amplitude through the multiplicative magnitude factor. As frequency increases, the amplitude decays, causing the phase contribution to be increasingly suppressed. As a result, the mid- and high-frequency phases—even those

[1] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, *Caltech-UCSD Birds-200-2011 Dataset*, Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[2] Wang et al., *Dataset Distillation with Neural Characteristic Function: A Minmax Perspective*, in *Proc. IEEE/ CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 25570–25580.

that still contain meaningful structural information—are overwhelmed and incorrectly treated as noise. To illustrate this behavior, we visualize phase differences across frequencies in Fig.1: low-frequency phase remains stable, the mid-frequency phase begins to drift, and the high-frequency phase becomes dominated by boundary noise. However, some med- and med-high-frequency phase components are still reliable and should not be discarded prematurely.

To restore these informative phase components, we extend the NCFM loss by adding an explicit phase constraint defined for each sampled frequency $\omega \sim p(t)$ that is decoupled from amplitude :

$$\begin{aligned} \mathrm{Chf}(t;f) &= \alpha \left(\left| \Phi_{f(x)}(t) - \Phi_{\hat{f}(x)}(t) \right|^2 \right) \\ &+ (1-\alpha) \left| \Phi_{f(x)}(t) \right| \left| \Phi_{\hat{f}(x)}(t) \right| \left(1 - \cos\left(a_{f(x)}(t) - a_{\hat{f}(x)}(t) \right) \right) \\ &+ \frac{\lambda_p}{1+\beta\|\omega\|^2} \left(\theta_{f(x)}(t) - \theta_{\hat{f}(x)}(t) \right)^2, \end{aligned} \tag{14}$$

This additional phase term explicitly extracts the remaining reliable phase information that is otherwise buried by amplitude attenuation in the original NCFM formulation. In particular, it allows mid-frequency phase—which is still semantically meaningful but degraded by the amplitude coupling—to be effectively preserved.

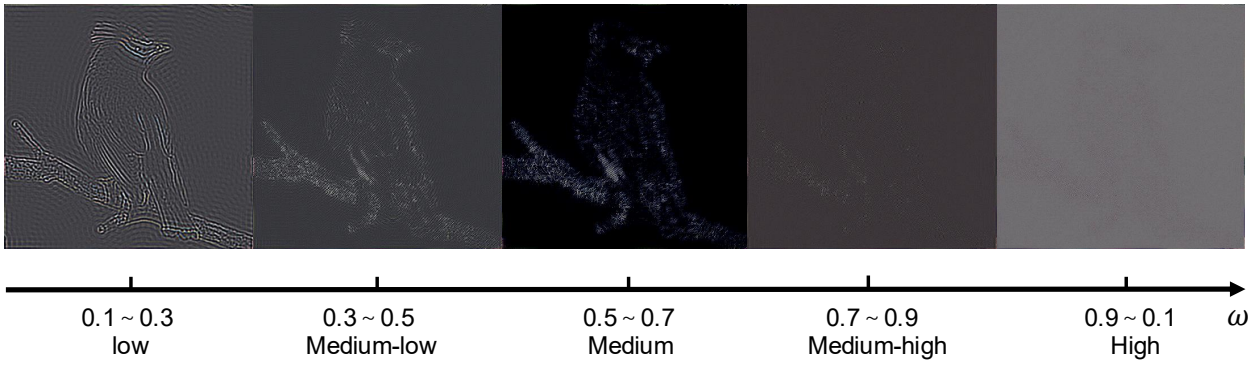


Figure 1: Phase information under different frequency band

Figure 2: Comparison between generated data with original loss function and PD-CFD loss.

We verify the effect of this decoupled phase constraint on downstream classification accuracy, and the results are illustrated in Table. 8 (see Fig. 2 for an example of the synthetic image). The results match our expectation: when $\lambda_p = 0.3$, the added phase regularization significantly strengthens high-frequency alignment, especially for the feather and edge regions in the CUB images.

A similar phenomenon is observed on the DTD and RESISC-45 datasets(illustrated in Fig. 4 Fig. 5 respectively). The DTD texture dataset contains a large amount of rapidly oscillating material patterns characterized by abrupt edge transitions, fine-scale boundaries, and dense local contour variations. Likewise, the RESISC-45 remote-sensing dataset

Table 8: Effect of phase regularization λ_p on accuracy.

λ_p	**Accuracy (%)**	**Improvement**
0.0	23.80	0
0.1	24.15	1.47%
0.2	25.20	5.88%
0.3	28.35	19.12%

exhibits substantial geometric complexity: land-cover boundaries, building edges, road networks, rooftop outlines, and other high-frequency landscape structures that vary sharply across classes. These discriminative details are dominated by high-order moment information in the frequency domain. Traditional metrics such as MSE or CE are inherently insensitive to such structural discontinuities—they primarily respond to low-order statistics and smooth variations, and therefore fail to capture the fine-grained distributional differences arising from high-frequency contours and edge transitions.

In contrast, our PD-CFD formulation introduced earlier can naturally recover these informative distributional discrepancies. By removing the amplitude-induced suppression of mid- and high-frequency phase, PD-CFD preserves the remaining reliable phase components that encode exactly these texture- and edge-based variations. Consequently, FAST maintains strong performance even on challenging datasets such as DTD and RESISC-45, where structural information is dominated by, high-frequency features and where competing baselines degrade substantially.

10 Experiments on Language Tasks

10.1 Setup

To further evaluate the generalization capability of FAST on LLM datasets, we conduct experiments on the Alpaca instruction-following dataset. Following our main setup, we adopt LLaMA2-7B as the backbone and apply LoRA-based fine-tuning using core-sets sampled at retention rates of 10%, 20%, and 30%. For evaluation, we report performance on the MMLU benchmark, which spans 57 diverse subjects and serves as a rigorous protocol for assessing broad knowledge and reasoning abilities. We compare FAST against the NMS baseline[3], the current SOTA method, under identical experimental settings.

10.2 Results

The MMLU results are shown in Fig. 6. Across all keep ratios, FAST consistently outperforms the NMS baseline. While NMS attains an average accuracy of approximately 33%, FAST increases the performance to 39%, corresponding to a relative improvement of about 18%. These results indicate that FAST is highly effective at selecting and preserving semantically informative instances. Moreover, the strong performance gains suggest that high-level semantic structure can be retained through spectral-graph-based distribution alignment alone, without explicit reliance on neural feature extractors. By leveraging geometric and frequency-domain signals, FAST preserves the underlying semantic neighborhood relations required for downstream reasoning. This demonstrates the robustness and broad generalization ability of FAST, even on tasks that rely heavily on semantic consistency.

[3]Boran Zhao et al., *NMS: Efficient Edge DNN Training via Near-Memory Sampling on Manifolds*, arXiv preprint arXiv:2508.02313, 2025.

Figure 3: CUB-200-2011 Bird Dataset

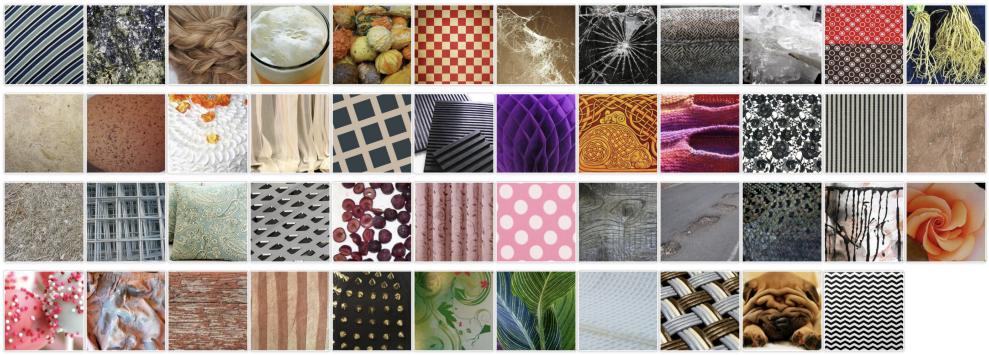
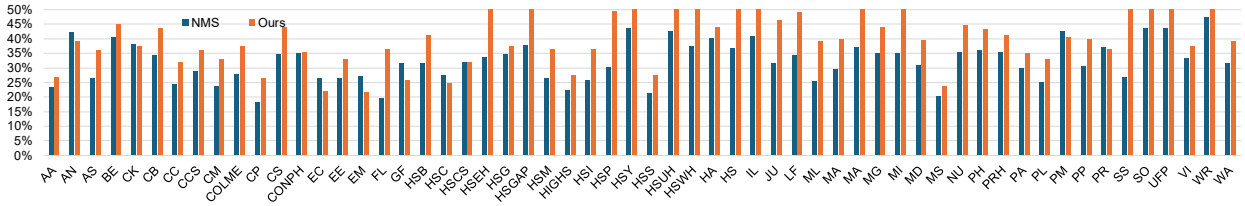
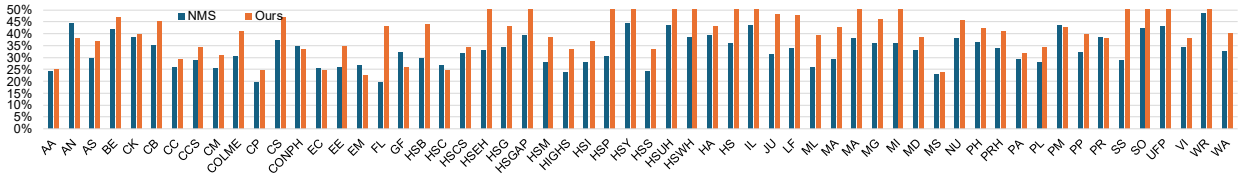


Figure 4: DTD Texture Dataset

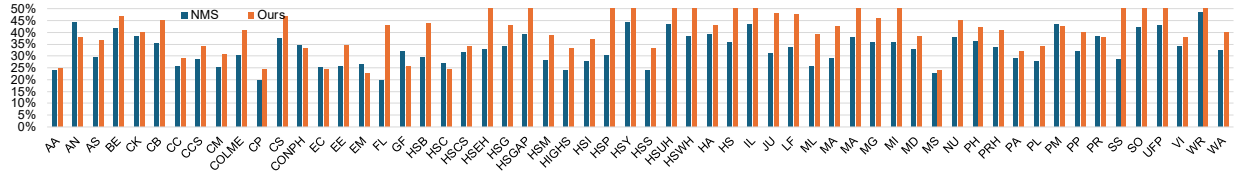
Figure 5: RESISC-45 Remote Sensing Dataset



(a) Keep ratio 10%



(b) Keep ratio 20%



(c) Keep ratio 30%

Figure 6: MMLU accuracy comparison at different keep ratios.