# Beyond Pairwise Supervision: Spectral Characteristic Matching for Data-Efficient Multimodal Alignment

**Xinchao Wang**[1] **Shuang Li**[1*] **Wei Chen**[1] **Jingxuan Kang**[2] **Fuzhen Zhuang**[1] **Deqin Wang**[1]

[1]Beihang University [2]Imperial College London

## Abstract

Multimodal alignment under scarce paired supervision aims to align independently pretrained unimodal encoders using limited image–text pairs, offering a practical alternative when large-scale joint pretraining is infeasible. Such scenarios call for lightweight objectives that go beyond local correspondences and capture distributional mismatch between heterogeneous embeddings. Existing methods often derive supervision from limited pairs through contrastive objectives or geometry-preserving regularization, encouraging instance-level correspondence or structural consistency but providing limited direct supervision on global cross-modal mismatch. This motivates a distribution-level objective that complements paired supervision by comparing image and text embedding distributions. In this paper, we propose **Spec**tral Characteristic Distribution **Align**ment (**SpecAlign**), a lightweight spectral alignment framework that reduces cross-modal mismatch by comparing image and text characteristic functions over learnable spectral queries. Since characteristic functions uniquely determine probability distributions, SpecAlign provides a principled signal for capturing multi-scale discrepancies beyond pointwise alignment. To obtain informative and stable queries, SpecAlign employs a Structured Direction–Radius Sampler (SDRS) that decouples direction and radius modeling. Extensive evaluations on transfer, retrieval, robustness, and representation analyses demonstrate consistent gains, validating spectral distribution-level supervision for data-efficient multimodal alignment.

## 1 Introduction

Multimodal models have achieved substantial progress, with shared image–text representations enabling effective transfer across classification, retrieval, and generation tasks [1, 2, 3]. Early visual-semantic embedding methods established joint representation spaces for recognition and retrieval [4, 5], while vision–language pretraining further advanced this paradigm through large-scale image–text supervision [1, 6, 7, 8, 9]. However, such success typically depends on massive paired corpora and substantial computation, as exemplified by CLIP's 400 million pairs [1], ALIGN's billion-scale noisy pairs [6], and subsequent scaling studies [7, 8, 10]. These requirements are difficult to meet in downstream domains where paired data are scarce, costly, or privacy-constrained [11, 12]. A common response is to adapt existing vision–language models with lightweight modules, including prompt-based adaptation [13, 14, 15, 16], adapters [17, 18, 19], and general parameter-efficient tuning methods [20, 21]. Yet these methods usually build on CLIP-style encoders that have already been jointly pretrained with paired image–text supervision, and therefore start from an approximately shared representation space.

This reliance on joint vision–language pretraining motivates a more data-efficient alternative: aligning independently pretrained vision and language encoders with limited paired supervision rather

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

than pretraining a multimodal model from scratch [22, 23]. This setting is more challenging because the two modalities may share compatible semantics while retaining different latent geometries and marginal distributions. Existing methods mainly rely on contrastive pairwise alignment or geometry-preserving regularization to preserve observed correspondences and modality-specific latent structure [22, 23, 24, 25]. Yet these objectives supervise only the observed cross-modal correspondences or their local geometry, leaving the marginal image and text distributions weakly constrained in the shared space. Consequently, local pairwise matching may align available pairs while still allowing a persistent global modality gap. Moreover, individual image–text pairs provide inherently underdetermined constraints, since the same caption can describe many visually diverse images and the same image can support multiple valid descriptions. Distribution-level matching aggregates such ambiguous local evidence into a population-level constraint, making alignment less sensitive to idiosyncratic pairs while better capturing modality-level discrepancies in location, scale, and structural variation.

This limitation motivates a distribution-level criterion for measuring cross-modal discrepancy beyond local pairwise neighborhoods. Characteristic functions provide such a criterion by representing distributions in the frequency domain, where image–text gaps can be probed across spectral scales. We therefore propose *Spectral Characteristic Distribution Alignment* (SpecAlign), which freezes pretrained encoders and learns lightweight projection heads to match modality distributions through empirical characteristic functions. Leveraging their distribution-determining property [26, 27], SpecAlign captures amplitude and phase discrepancies as stable population-level constraints complementary to pairwise supervision. We further introduce a Structured Direction–Radius Sampler (SDRS) to avoid limited fixed-probe coverage [28, 29, 30] by learning mismatch-aware frequency directions with controlled magnitudes.

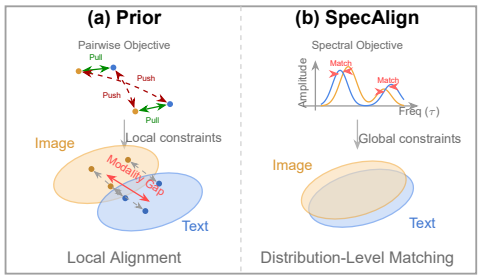



Figure 1: Schematic comparison under limited paired supervision. Prior methods impose pairwise local constraints, which may align observed pairs while still leaving a modality gap between image and text distributions. SpecAlign instead performs spectral distribution-level matching, imposing global constraints that encourage alignment between image and text distributions.

SpecAlign improves zero-shot classification, cross-modal retrieval, distribution-shift robustness, and optimization stability, while spectral analyses confirm reduced cross-modal frequency mismatch. Our contributions are threefold: (i) we recast low-resource multimodal alignment as shared-space distribution matching and instantiate it with a characteristic-function objective; (ii) we introduce SDRS for stable frequency-domain probing, supported by theoretical analysis of how spectral discrepancies reflect cross-modal distribution mismatch; and (iii) we validate the theory and method through extensive experiments across transfer, retrieval, robustness and spectral analyses.

## 2 Related Work

### 2.1 Low-Resource and Parameter-Efficient Multimodal Alignment

Low-resource multimodal alignment aims to connect strong pretrained vision and language encoders using limited paired data [23, 24, 31, 32, 33]. A common strategy is to freeze the encoders and learn lightweight projectors [24, 31], prompt-based modules [13, 14, 34, 35], adapters or low-rank updates [17, 18, 20, 21], or regularized/test-time prompt variants [16, 36, 37]. Structure-guided and parameter-efficient variants further preserve unimodal geometry through visual prompts, task prompts, distillation, or neighborhood regularization [15, 19, 22, 23, 25, 38, 39].

Overall, these approaches mainly retain local relations or modality-specific structure in the shared embedding space. Complementary work studies visual-semantic embedding objectives [5], richer image-text representation geometries [40], and compositional cross-modal relations [41]. In contrast, SpecAlign uses characteristic-function-based distribution matching to provide direct distribution-level supervision, allowing the aligned space to capture global, multi-scale modality discrepancies that are not fully constrained by local geometry preservation under scarce paired supervision.

### 2.2 Pretrained Encoders and Vision-Language Foundation Models

Low-resource multimodal alignment commonly builds on strong pretrained unimodal encoders and focuses on aligning their frozen representations [23, 24, 31, 32]. This setting builds on transformer-based language modeling [42, 43] and vision backbones including ViT [44] and the DINO family [45, 46, 47]. It also relates to large-scale vision-language foundation models that combine visual encoders with language supervision [8, 9, 10, 48, 49, 50] or language-model interfaces and cross-modal alignment frameworks [2, 3, 32, 51, 52, 53, 54]. Recent studies of multimodal representation geometry and transfer further analyze partial concept overlap and implicit alignment across modalities [55, 56], while broader multimodal alignment frameworks continue to extend joint-embedding approaches [33]. Recent analysis of pretrained vision backbones further suggests that the strongest transferable visual embeddings may arise at intermediate rather than output layers, which supports layer-aware use of frozen encoders in alignment settings [57]. Rather than training such models from scratch, our goal is to make their frozen representations more compatible in data-scarce scenarios.

### 2.3 Frequency-Domain Modeling for Distribution Alignment

Frequency-domain representations [28] and characteristic-function-based perspectives [26, 27, 58] offer a principled way to characterize global and multi-scale structure in learned representations. Fourier features [28, 59], learned two-sample tests [29], recent spectral parameterizations [60, 61], and characteristic-function-based objectives [30] have shown the value of analyzing distributions through their spectral signatures, where discrepancies in energy, phase, and higher-order structure can be made explicit. Despite these advances, frequency-domain modeling has rarely been used as a direct objective for low-resource multimodal alignment. SpecAlign builds on this line of work by equipping characteristic-function matching with SDRS, which learns mismatch-aware frequency directions and controls probe magnitudes through a shared radius, thereby making spectral distribution matching more informative and stable for low-resource multimodal alignment.

## 3 Preliminaries

### 3.1 Low-Resource Multimodal Alignment Setup

Given a paired dataset $\mathcal{D} = \{(\mathbf{x}_v^i, \mathbf{x}_t^i)\}_{i=1}^N \subseteq \mathcal{X}_v \times \mathcal{X}_t$ with limited supervision, we consider the low-resource multimodal alignment problem. Let $f_v : \mathcal{X}_v \to \mathbb{R}^{d_v}$ and $f_t : \mathcal{X}_t \to \mathbb{R}^{d_t}$ denote frozen pretrained visual and textual encoders, respectively. For the $i$-th paired sample, they produce modality-specific features $\mathbf{h}_v^i = f_v(\mathbf{x}_v^i)$ and $\mathbf{h}_t^i = f_t(\mathbf{x}_t^i)$.

Our goal is to learn lightweight mappings $g_v : \mathbb{R}^{d_v} \to \mathcal{A}$ and $g_t : \mathbb{R}^{d_t} \to \mathcal{A}$ that project the frozen features into a shared alignment space $\mathcal{A} \subseteq \mathbb{R}^k$, as commonly done in low-resource cross-modal adaptation settings [24, 31]. The aligned representations are then defined as $\mathbf{z}_v^i = g_v(\mathbf{h}_v^i)$ and $\mathbf{z}_t^i = g_t(\mathbf{h}_t^i)$.

Before characteristic-function matching, we apply sample-wise $\ell_2$ normalization to the aligned representations:

$$\bar{\mathbf{z}}_u^i = \frac{\mathbf{z}_u^i}{\|\mathbf{z}_u^i\|_2 + \epsilon}, \quad u \in \{v, t\}, \tag{1}$$

where $\epsilon > 0$ ensures numerical stability. This normalization largely removes dependence on raw embedding magnitudes during spectral comparison.

Since only $g_v$ and $g_t$ are optimized, the available supervision is concentrated on cross-modal adaptation in the shared space. Let $Q_v$ and $Q_t$ denote the feature distributions induced by the frozen encoders $f_v$ and $f_t$, respectively. The normalized aligned distributions are denoted by $P_v = (n \circ g_v)_* Q_v$ and $P_t = (n \circ g_t)_* Q_t$, where $n(\mathbf{z}) = \mathbf{z}/(\|\mathbf{z}\|_2 + \epsilon)$ and $(\cdot)_*$ denotes the pushforward operation [23, 32].

### 3.2 Characteristic Functions

To compare image and text embedding distributions in the shared space, we use characteristic functions as Fourier-domain representations of probability distributions [26]. For a normalized random

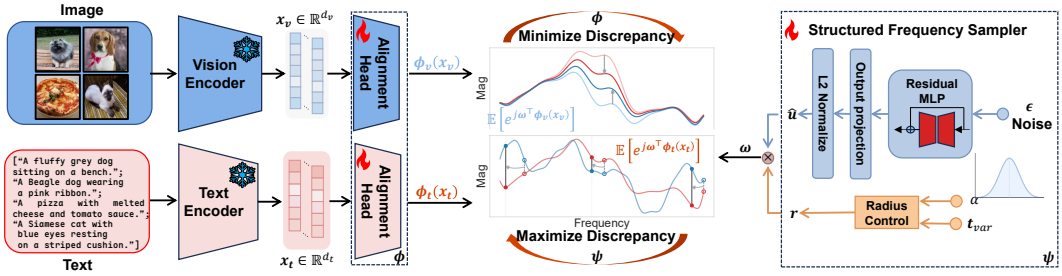



Figure 2: Overview of SpecAlign. Frozen image and text encoders are mapped into a shared embedding space by lightweight alignment heads. SpecAlign complements paired contrastive learning with characteristic-function matching, where empirical spectral amplitude and phase discrepancies are minimized between modalities. A structured frequency sampler adversarially generates informative probes by learning frequency directions while controlling their radii, yielding stable and mismatch-aware distribution alignment.

embedding $\bar{\mathbf{z}} \sim P$ and frequency vector $\boldsymbol{\tau} \in \mathbb{R}^k$, the characteristic function is

$$\phi_P(\boldsymbol{\tau}) = \mathbb{E}_{\bar{\mathbf{z}} \sim P}\left[e^{\mathrm{i}\boldsymbol{\tau}^\top \bar{\mathbf{z}}}\right], \tag{2}$$

where i is the imaginary unit. Since characteristic functions uniquely determine probability distributions [26], their matching provides a principled criterion for distribution-level alignment.

Let $P_v$ and $P_t$ denote the normalized image and text embedding distributions, with characteristic functions $\phi_v$ and $\phi_t$. Given a probing distribution $F_\psi$, we define

$$\mathcal{C}_\psi(P_v, P_t) = \mathbb{E}_{\boldsymbol{\tau} \sim F_\psi}\left[\left|\phi_v(\boldsymbol{\tau}) - \phi_t(\boldsymbol{\tau})\right|^2\right]. \tag{3}$$

In practice, for a minibatch of normalized features $\{\bar{\mathbf{z}}^i\}_{i=1}^B$, we use the differentiable empirical estimate $\hat{\phi}_P(\boldsymbol{\tau}) = B^{-1}\sum_{i=1}^B e^{\mathrm{i}\boldsymbol{\tau}^\top \bar{\mathbf{z}}^i}$, which enables stochastic optimization. Varying $\boldsymbol{\tau}$ compares distributions across spectral scales, from coarse low-frequency mismatch to finer high-frequency discrepancies.

**Proposition 3.1** (Distribution-identifying spectral discrepancy)**.** *Assume that $\Omega \subset \mathrm{supp}(F_\psi)$ for some non-empty open set $\Omega \subset \mathbb{R}^k$, and that $P_v$ and $P_t$ admit moment generating functions near the origin. If $\mathcal{C}_\psi(P_v, P_t) = 0$, then $P_v = P_t$.*

**Theoretical scope and practical interpretation.** The sample-wise normalization in Eq. (1) ensures $\|\bar{\mathbf{z}}\|_2 \leq 1$. Consequently, for any $\mathbf{s} \in \mathbb{R}^k$, the moment generating function satisfies $\mathbb{E}[e^{\mathbf{s}^\top \bar{\mathbf{z}}}] \leq e^{\|\mathbf{s}\|_2} < \infty$. Thus, the normalized feature distributions satisfy the analyticity condition required by Proposition 3.1. However, the open-set spectral coverage assumption is a population-level condition that is not guaranteed by finite, radius-constrained SDRS probes. Therefore, the practical objective reduces distributional discrepancies over the learned spectral probes rather than certifying exact equality of the full distributions.

## 4 Method

SpecAlign freezes pretrained image and text encoders and learns lightweight projection heads into a shared embedding space. As illustrated in Figure 2, it combines paired contrastive supervision with empirical characteristic-function matching, where SDRS learns scale-controlled frequency probes to expose cross-modal distribution mismatch. We next introduce the spectral objective, its interpretations, and its connection to downstream score stability.

For paired inputs, we write $\mathbf{z}_v^i = g_v(f_v(\mathbf{x}_v^i))$ and $\mathbf{z}_t^i = g_t(f_t(\mathbf{x}_t^i))$. Let $P_v$ and $P_t$ denote the induced shared-space modality distributions, and let $\{\boldsymbol{\tau}_m\}_{m=1}^M$ denote the sampled frequency probes. Unless otherwise specified, $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity.

### 4.1 Characteristic-Function-Based Alignment for Different Modalities

SpecAlign formulates multimodal alignment as matching the characteristic functions of $P_v$ and $P_t$ in the shared embedding space. Given probes $\boldsymbol{\tau}_m \sim F_\psi$, where $F_\psi$ is specified in Section 4.2, we

estimate the characteristic function of modality $u \in \{v, t\}$ by

$$\hat{\phi}_u(\boldsymbol{\tau}_m) = \frac{1}{B} \sum_{i=1}^{B} e^{\mathrm{i}\boldsymbol{\tau}_m^\top \mathbf{z}_u^i}, \tag{4}$$

with population counterpart $\phi_u(\boldsymbol{\tau}) = \mathbb{E}_{\mathbf{z}\sim P_u}[e^{\mathrm{i}\boldsymbol{\tau}^\top \mathbf{z}}]$. The pointwise and population spectral discrepancies are

$$d_{\mathrm{CF}}(\boldsymbol{\tau}) = \left|\phi_v(\boldsymbol{\tau}) - \phi_t(\boldsymbol{\tau})\right|^2, \tag{5}$$

$$\mathcal{C}_\psi(P_v, P_t) = \int d_{\mathrm{CF}}(\boldsymbol{\tau})\, dF_\psi(\boldsymbol{\tau}), \tag{6}$$

which is approximated in minibatches as

$$\widehat{\mathcal{C}}_\psi = \frac{1}{M} \sum_{m=1}^{M} \left|\hat{\phi}_v(\boldsymbol{\tau}_m) - \hat{\phi}_t(\boldsymbol{\tau}_m)\right|^2. \tag{7}$$

To separate spectral energy mismatch from structural offset, we write $\phi_u(\boldsymbol{\tau}) = \rho_u(\boldsymbol{\tau})e^{\mathrm{i}a_u(\boldsymbol{\tau})}$. Substitution into Eq. (5) yields

$$d_{\mathrm{CF}}(\boldsymbol{\tau}) = \big(\rho_v(\boldsymbol{\tau}) - \rho_t(\boldsymbol{\tau})\big)^2 + 2\rho_v(\boldsymbol{\tau})\rho_t(\boldsymbol{\tau})\big(1 - \cos(a_v(\boldsymbol{\tau}) - a_t(\boldsymbol{\tau}))\big). \tag{8}$$

The first term captures amplitude discrepancy, while the second captures phase mismatch. SpecAlign therefore uses the amplitude-phase alignment loss

$$\mathcal{L}_{\mathrm{SpecAlign}} = \frac{1}{M} \sum_{m=1}^{M} \sqrt{\alpha_{\mathrm{cf}} \mathcal{L}_{\mathrm{amp}}^{(m)} + \beta_{\mathrm{cf}} \mathcal{L}_{\mathrm{pha}}^{(m)}}, \tag{9}$$

where

$$\mathcal{L}_{\mathrm{amp}}^{(m)} = \Big(|\hat{\phi}_v(\boldsymbol{\tau}_m)| - |\hat{\phi}_t(\boldsymbol{\tau}_m)|\Big)^2, \tag{10}$$

$$\mathcal{L}_{\mathrm{pha}}^{(m)} = 2|\hat{\phi}_v(\boldsymbol{\tau}_m)||\hat{\phi}_t(\boldsymbol{\tau}_m)|\left(1 - \cos\left(\hat{a}_v(\boldsymbol{\tau}_m) - \hat{a}_t(\boldsymbol{\tau}_m)\right)\right), \tag{11}$$

and $\hat{a}_u(\boldsymbol{\tau}_m) = \arg(\hat{\phi}_u(\boldsymbol{\tau}_m))$.

We combine SpecAlign with the bidirectional paired contrastive objective. Let $q_{ij}^{vt} = \exp(\mathrm{sim}(\mathbf{z}_v^i, \mathbf{z}_t^j)/\kappa)$ and $q_{ij}^{tv} = \exp(\mathrm{sim}(\mathbf{z}_t^i, \mathbf{z}_v^j)/\kappa)$. Then

$$\mathcal{L}_{\mathrm{CLIP}} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{q_{ii}^{vt}}{\sum_{j=1}^{B} q_{ij}^{vt}} + \log \frac{q_{ii}^{tv}}{\sum_{j=1}^{B} q_{ij}^{tv}}\right], \tag{12}$$

and the full objective is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{CLIP}} + \lambda_{\mathrm{SpecAlign}} \mathcal{L}_{\mathrm{SpecAlign}}. \tag{13}$$

**Proposition 4.1** (Low-frequency interpretation of SpecAlign)**.** *Assume the relevant third moments of $P_v$ and $P_t$ are finite. Let $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_v - \boldsymbol{\mu}_t$ and $\Delta\Sigma = \Sigma_v - \Sigma_t$. For small $\|\boldsymbol{\tau}\|$,*

$$d_{\mathrm{CF}}(\boldsymbol{\tau}) = \big(\boldsymbol{\tau}^\top \Delta\boldsymbol{\mu}\big)^2 + \frac{1}{4}\big(\boldsymbol{\tau}^\top \Delta\Sigma\, \boldsymbol{\tau}\big)^2 + O(\|\boldsymbol{\tau}\|^4). \tag{14}$$

Proposition 4.1 shows that low frequencies capture mean and covariance gaps, while larger frequencies reveal finer distributional discrepancies. The derivation and a translation example are provided in Appendix A. Since different frequencies expose different levels of distributional mismatch, SpecAlign requires probes that are both informative and stable during training. We next describe a structured sampler for this purpose.

### 4.2 Structured Direction–Radius Parameterization for Frequency Sampling

The effectiveness of Eq. (6) depends on how $F_\psi$ places mass in frequency space. Learnable probes can emphasize informative spectral regions [28, 29, 30, 59, 60, 61]. However, generic MLP samplers use an unconstrained mapping that entangles direction selection with norm scaling. Under adversarial training, this entanglement can create a radial shortcut: the sampler may enlarge frequency norms to amplify the discrepancy, rather than identify meaningful directions of cross-modal mismatch. To avoid this, SpecAlign parameterizes $F_\psi$ with a Structured Direction–Radius Sampler (SDRS):

$$s_\psi(\boldsymbol{\epsilon}) = r_0 \cdot \mathrm{clip}(\exp(\alpha), r_{\min}, r_{\max}) \frac{W_o h_\psi(\boldsymbol{\epsilon}) + b_o}{\|W_o h_\psi(\boldsymbol{\epsilon}) + b_o\|_2}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I). \tag{15}$$

Here $h_\psi$ is a learnable MLP, $W_o$ and $b_o$ are output projection parameters, $\alpha$ is a learnable log-radius parameter, and $r_0, r_{\min}, r_{\max}$ determine the controlled radius range. SDRS therefore separates sample-specific direction learning from shared radius control.

**Lemma 4.2** (Radius-controlled directional probing)**.** *Let $\mathbf{u}_m = W_o h_\psi(\boldsymbol{\epsilon}_m) + b_o$, $\hat{\mathbf{u}}_m = \mathbf{u}_m / \|\mathbf{u}_m\|_2$, and $r = r_0 \cdot \mathrm{clip}(\exp(\alpha), r_{\min}, r_{\max})$, so that $\boldsymbol{\tau}_m = s_\psi(\boldsymbol{\epsilon}_m) = r\hat{\mathbf{u}}_m$. Then, in one forward pass, for every $m = 1, \ldots, M$, $\|\boldsymbol{\tau}_m\|_2 = r$, $\mathrm{Var}_{m=1,\ldots,M}\big(\|\boldsymbol{\tau}_m\|_2\big) = 0$, and $r_0\, r_{\min} \le \|\boldsymbol{\tau}_m\|_2 \le r_0\, r_{\max}$. Moreover, for $\boldsymbol{\tau} = r\hat{\mathbf{u}}$ with $\hat{\mathbf{u}} = \mathbf{u}/\|\mathbf{u}\|_2$ and $\mathbf{u} \neq \mathbf{0}$,*

$$\frac{\partial \boldsymbol{\tau}}{\partial \mathbf{u}} = \frac{r}{\|\mathbf{u}\|_2} \left(I - \hat{\mathbf{u}}\hat{\mathbf{u}}^\top\right), \qquad \frac{\partial \boldsymbol{\tau}}{\partial \mathbf{u}} (c\,\hat{\mathbf{u}}) = \mathbf{0}. \tag{16}$$

*Here $I$ denotes the identity matrix and $c$ any scalar. If the aligned embeddings are cosine-normalized, then $|\boldsymbol{\tau}^\top \mathbf{z}| \le r_0\, r_{\max}$.*

Lemma 4.2 formalizes the optimization effect of SDRS. The shared bounded radius removes probe-wise norm variation, preventing the adversarial sampler from increasing the discrepancy through uncontrolled frequency magnitudes. Meanwhile, the tangent-space Jacobian eliminates first-order sensitivity to sample-specific radial perturbations, so the sampler is driven to search over directions on a controlled hypersphere. Thus, SDRS turns adversarial frequency learning into a stable directional search for cross-modal mismatch.

During training, the alignment heads $\theta_v, \theta_t$ minimize $\mathcal{L}_{\mathrm{total}}$, while the sampler parameter $\psi$ is updated adversarially to maximize $\mathcal{L}_{\mathrm{SpecAlign}}$. Algorithm 1 summarizes one training iteration, and proof details are provided in Appendix A.

### 4.3 Conditional Distribution and Downstream Interpretation

The low-frequency analysis explains what SpecAlign matches at small $\|\boldsymbol{\tau}\|$. We next relate spectral discrepancy to downstream score transfer. The following result is a conditional stability statement: it applies when the probing density $p_\psi$ covers the Fourier content of the downstream score.

**Theorem 4.3** (Conditional spectral control of downstream score drift)**.** *Under the regularity and coverage assumptions stated in Appendix A, any compatible score $h : \mathbb{R}^k \to \mathbb{R}$ with Fourier transform $\widehat{h}$ satisfies*

$$|\mathbb{E}_{P_v}[h(\mathbf{z})] - \mathbb{E}_{P_t}[h(\mathbf{z})]| \le \left( \int_{\mathbb{R}^k} \frac{|\widehat{h}(\boldsymbol{\tau})|^2}{p_\psi(\boldsymbol{\tau})}\, d\boldsymbol{\tau} \right)^{1/2} \sqrt{\mathcal{C}_\psi(P_v, P_t)}. \tag{17}$$

Thus, reducing $\mathcal{C}_\psi$ controls downstream score drift only for functions whose Fourier energy is sufficiently covered by $p_\psi$; the bound is not by itself a finite-sample accuracy guarantee.

**Corollary 4.4** (Conditional expected margin preservation for zero-shot classification)**.** *For class score $s_c(\mathbf{z}) = \langle \mathbf{z}, \mathbf{p}_c \rangle$ with class prototype $\mathbf{p}_c$ and ground-truth class $y$, suppose*

$$\begin{aligned} \bar{\gamma}_v &:= \mathbb{E}_{P_v}[s_y(\mathbf{z})] - \max_{c \neq y} \mathbb{E}_{P_v}[s_c(\mathbf{z})] \ge \gamma_{\min} > 0, \\ |\mathbb{E}_{P_v}[s_c(\mathbf{z})] - \mathbb{E}_{P_t}[s_c(\mathbf{z})]| &\le B_c \sqrt{\mathcal{C}_\psi(P_v, P_t)} \quad \textit{for all } c. \end{aligned} \tag{18}$$

*Let $B = \max_c B_c$. Then $2B\sqrt{\mathcal{C}_\psi(P_v, P_t)} < \gamma_{\min}$ implies $\mathbb{E}_{P_t}[s_y(\mathbf{z})] > \max_{c \neq y} \mathbb{E}_{P_t}[s_c(\mathbf{z})]$.*

Corollary 4.4 gives an expectation-level explanation of when distribution matching can preserve class-score ordering after cross-modal transfer. Together, Theorem 4.3 and Corollary 4.4 provide a conditional explanation of score stability, with proofs deferred to Appendix A.

Table 1: Zero-shot classification results under the shared frozen-encoder low-resource alignment protocol. We report *last* for final-layer features and *Similar* for the most similar cross-modal layer pair. Best results are in **bold**, second-best results are underlined, and Average is computed column-wise over all classification datasets.

| Zero-shot Classification Accuracy (%) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | **LilT** [24] (ICLR'23) *last* | **LilT** [24] (ICLR'23) *Simi* | **MLP** [22] (CVPR'25) *last* | **MLP** [22] (CVPR'25) *Simi* | **CSA** [25] (ICLR'25) *last* | **CSA** [25] (ICLR'25) *Simi* | **STRU** [23] (NeurIPS'25) *on LilT* | **STRU** [23] (NeurIPS'25) *on MLP* | **STRU** [23] (NeurIPS'25) *on CSA* | **SpecAlign** (Ours) *last* | **SpecAlign** (Ours) *Simi* |
| STL10 | 75.6 | 79.7 | 76.6 | 84.0 | 77.9 | 80.0 | 92.6 | 92.7 | 91.7 | 92.0 | **93.6** |
| CIFAR10 | 85.5 | 89.0 | 79.2 | 81.5 | 78.5 | 80.8 | 96.3 | 96.3 | 97.2 | 96.3 | **97.3** |
| CIFAR100 | 34.0 | 33.1 | 35.3 | 34.5 | 47.4 | 47.4 | 51.3 | 52.1 | **56.4** | 51.5 | 52.1 |
| EuroSAT | 19.9 | 22.4 | 25.4 | 22.2 | 26.4 | 22.5 | 29.2 | 25.0 | 28.4 | 35.6 | **37.7** |
| GTSRB | 3.4 | 4.0 | 7.8 | 7.4 | 7.0 | 6.0 | 7.8 | 7.9 | 4.9 | 8.9 | **11.4** |
| Pets | 7.0 | 4.9 | 5.3 | 6.1 | 14.4 | 14.9 | 13.2 | 15.3 | 17.0 | **18.3** | 17.4 |
| Food101 | 14.8 | 14.6 | 15.6 | 17.1 | 29.3 | 28.0 | 30.6 | 30.5 | 28.6 | 30.7 | **32.6** |
| Caltech101 | 37.9 | 39.5 | 38.2 | 38.8 | 31.4 | 33.6 | 56.0 | 56.0 | **61.8** | 55.0 | 58.5 |
| ImageNet | 9.9 | 10.5 | 10.6 | 11.4 | 23.2 | 23.3 | 24.5 | 25.1 | **26.8** | 24.1 | 24.9 |
| UCF101 | 21.5 | 20.4 | 27.7 | 25.7 | 35.5 | 34.4 | 41.8 | 42.3 | 44.0 | 41.9 | **44.1** |
| KITTI Distance | 26.8 | 29.5 | 24.1 | 25.4 | 27.1 | 22.3 | 28.5 | 32.7 | 31.6 | **34.3** | 33.9 |
| DTD | 9.0 | 8.2 | 8.4 | 9.4 | 12.6 | 13.0 | 18.4 | 18.7 | 14.8 | 18.9 | **19.2** |
| **Average** | 28.8 | 29.7 | 29.5 | 30.3 | 34.2 | 33.9 | 40.9 | 41.2 | 41.9 | 42.3 | **43.5** |

# 5 Experiments

## 5.1 Experimental Setup

**Datasets.** We train the alignment modules using paired image–text examples from the MS-COCO 2014 training split, following the low-resource setting where pretrained unimodal encoders are kept frozen and only lightweight alignment components are optimized. To assess the transferability of the learned shared representation, we evaluate zero-shot classification on a diverse set of recognition benchmarks, including STL10 [62], CIFAR10 and CIFAR100 [63], EuroSAT [64], GTSRB [65], Oxford-IIIT Pets [66], Food101 [67], Caltech101 [68], ImageNet [69], UCF101 [70], KITTI Distance [71], and DTD [72]. These datasets cover natural images, fine-grained recognition, textures, scenes, traffic signs, satellite imagery, and action-related categories, providing a broad testbed for zero-shot transfer. We further evaluate robustness on ImageNet-A [73], ImageNet-R [74], ImageNet-Sketch [75], and ImageNet-V2 [76], and assess cross-modal retrieval on Flickr30K [77] and MS-COCO [78].

**Alignment setting and baselines.** Following the evaluation protocol of STRUCTURE [23], we use frozen DINOv2 and RoBERTa encoders with lightweight gated low-rank alignment heads, and train SpecAlign under the same frozen-encoder, paired-data, training-epoch, and evaluation settings. We report a *last* setting using final-layer features and a *Similar* setting using the most similar cross-modal layer pair from the baseline protocol; this is also consistent with evidence that strong visual embeddings need not arise at the network output [23, 57]. We compare against LilT [24], MLP [22], CSA [25], and STRUCTURE [23]. For baselines reported in STRUCTURE [23], we follow their published numbers to maintain consistency with prior low-resource alignment evaluations. Detailed backbone, optimization, and hyperparameter settings are given in Appendix C. For evaluation metrics, we report average top-1 accuracy for zero-shot classification and Recall@K for image-to-text and text-to-image retrieval, with emphasis on R@1.

## 5.2 Main Results

We first examine whether distribution-level spectral supervision improves low-resource alignment under the frozen-encoder protocol used by prior methods. Under matched encoders, alignment heads, training pairs, and training budgets, SpecAlign achieves the best average zero-shot classification accuracy and the strongest overall retrieval and robustness results among the compared methods. Tables 1, 2, and 3 summarize the results on zero-shot classification, cross-modal retrieval, and shifted-ImageNet robustness, respectively.

These tasks probe complementary properties of the shared space: classification depends on the calibration of text prototypes against image features, retrieval directly evaluates cross-modal neighbor-

Table 2: Zero-shot cross-modal retrieval. We report image-to-text and text-to-image Recall@1 on Flickr30K and COCO. Best results are in **bold** and second-best results are underlined.

| Method | Flickr30K | | COCO | |
|---|---|---|---|---|
| | I2T@1 | T2I@1 | I2T@1 | T2I@1 |
| LilT *(last)* | 32.5 | 22.1 | 6.9 | 3.6 |
| LilT *(Simi)* | 35.3 | 24.0 | 7.1 | 4.0 |
| MLP *(last)* | 31.6 | 20.3 | 6.0 | 2.7 |
| MLP *(Simi)* | 36.4 | 25.0 | 7.1 | 3.8 |
| CSA *(last)* | 47.0 | 38.3 | 9.2 | 8.0 |
| CSA *(Simi)* | 48.6 | 39.0 | 9.6 | 8.1 |
| STRUCTURE *(on LilT)* | 65.8 | 53.7 | 18.2 | 13.3 |
| STRUCTURE *(on MLP)* | 65.9 | 53.8 | 18.3 | 13.4 |
| STRUCTURE *(on CSA)* | 56.1 | 43.1 | 10.3 | 8.1 |
| **SpecAlign** *(last)* | 66.3 | 53.8 | 18.1 | 13.2 |
| **SpecAlign** *(Simi)* | **67.4** | **54.1** | **19.4** | **13.8** |

Table 3: Distribution-shift robustness. We report zero-shot classification accuracy on four shifted ImageNet benchmarks. Best results are in **bold** and second-best results are underlined.

| Method | ImageNet Distribution Shifts | | | | Avg. |
|---|---|---|---|---|---|
| | -A | -R | -V2 | -Sketch | |
| LilT *(last)* | 19.0 | 28.6 | 19.0 | 17.3 | 21.1 |
| LilT *(Simi)* | 19.7 | 29.4 | 20.3 | 17.6 | 21.8 |
| MLP *(last)* | 17.8 | 28.8 | 18.8 | 17.7 | 20.8 |
| MLP *(Simi)* | 17.0 | 29.9 | 18.9 | 17.5 | 20.8 |
| CSA *(last)* | 21.8 | 32.9 | 22.0 | 15.9 | 23.2 |
| CSA *(Simi)* | 23.5 | 32.1 | 20.6 | 16.7 | 23.2 |
| STRUCTURE *(on LilT)* | 24.0 | 33.0 | 20.7 | 18.7 | 24.1 |
| STRUCTURE *(on MLP)* | 22.6 | 32.5 | 22.2 | 20.7 | 24.5 |
| STRUCTURE *(on CSA)* | 24.3 | 33.2 | 21.9 | 21.3 | 25.2 |
| **SpecAlign** *(last)* | 24.8 | **34.4** | 22.1 | **21.6** | **25.7** |
| **SpecAlign** *(Simi)* | **24.9** | 33.7 | **22.3** | 21.4 | 25.6 |

hood ranking, and shifted-ImageNet benchmarks test whether the learned alignment remains useful under visual distribution shift. The gains are broad but not uniform across datasets. SpecAlign obtains the best classification average, with broad improvements across fine-grained, scene, texture, and traffic-oriented benchmarks. Although the gains vary across datasets, the overall trend indicates that spectral distribution matching improves the global compatibility between image and text representations, rather than benefiting only a narrow benchmark subset.

Retrieval results provide a complementary perspective. SpecAlign improves both image-to-text and text-to-image Recall@1 on Flickr30K and COCO, suggesting that the learned shared space benefits not only class-level prototype matching but also instance-level cross-modal ranking. On shifted ImageNet benchmarks, SpecAlign further achieves the best average robustness score, indicating that reducing modality-level distribution mismatch helps preserve useful decision structure under visual distribution shift. Under the frozen-encoder, paired-data, training-epoch, and evaluation protocol adopted from STRUCTURE, these results primarily reflect differences in the alignment objective rather than changes in backbone capacity or training data.

### 5.3 Analysis I: Sampler Architecture and Optimization Stability

To isolate the effect of sampler parameterization, we compare SDRS with SampleNet, a standard MLP-based frequency sampler, while keeping the frozen encoders, alignment heads, paired training data, loss weights, and optimization schedule unchanged. As shown in Table 4, SDRS improves zero-shot classification on all evaluated datasets, with gains ranging from $+1.8$ points on Food101 to $+4.4$ points on CIFAR10 and an average improvement of about $+3.0$ points across the five reported benchmarks.

Table 4: Ablation on sampler architecture for zero-shot classification accuracy.

| Dataset | SampleNet | SDRS | $\Delta$ |
|---|---|---|---|
| STL10 | 90.2 | **93.5** | +3.3 |
| CIFAR10 | 92.9 | **97.3** | +4.4 |
| CIFAR100 | 49.7 | **52.0** | +2.3 |
| Food101 | 29.5 | **31.3** | +1.8 |
| Caltech101 | 54.5 | **57.8** | +3.3 |

These gains indicate that the structured sampler contributes beyond the spectral objective itself, yet accuracy alone does not explain why the parameterization helps. We therefore examine the sampled-frequency norms and optimization dynamics. In characteristic-function matching, sampled frequencies determine where image and text embedding distributions are compared; a direct noise-to-frequency MLP may entangle directional search with uncontrolled norm variation, making the discrepancy sensitive to unstable frequency scales. SDRS instead decouples these factors by learning informative directions while using a shared bounded radius to keep the probing scale controlled. Figures 3 and 4 support this interpretation: compared with SampleNet, SDRS produces more concentrated sampled-frequency norms and smoother training and validation loss trajectories. Together with the accuracy gains in Table 4, these diagnostics suggest that SDRS improves performance not merely by increasing sampler capacity, but by turning frequency probing into a more stable direction-search problem, consistent with Lemma 4.2.

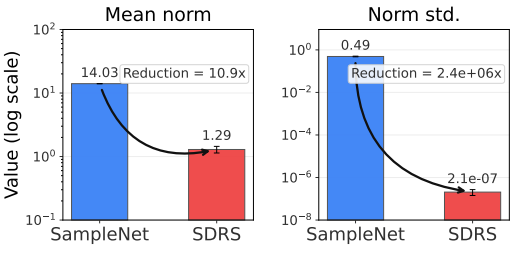



Figure 3: Norm statistics of the sampler. We report the mean and standard deviation of sampler norms for SDRS and SampleNet.

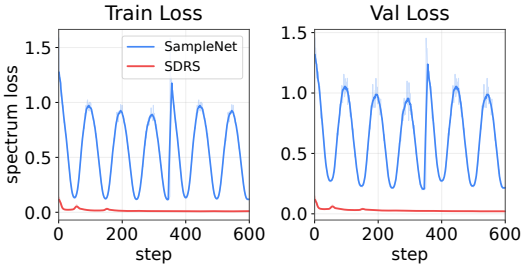



Figure 4: Optimization dynamics of the sampler. We show the training and validation loss curves for SDRS and SampleNet, normalized by their respective epoch-0 loss values.

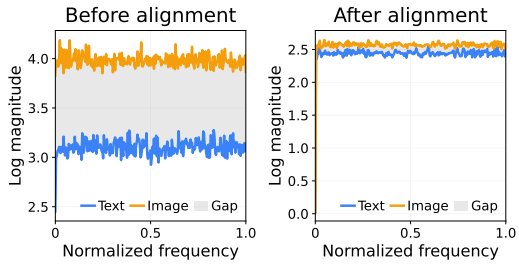



Figure 5: Raw spectral magnitude profiles on the COCO validation split before and after alignment. SpecAlign reduces the observed magnitude gap over the evaluated frequencies, without implying complete distributional equality.

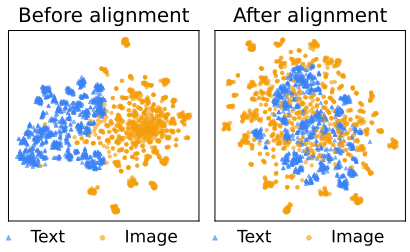



Figure 6: Shared-space t-SNE on the COCO validation split before and after alignment. Image and text embeddings exhibit reduced modality separation after SpecAlign.

### 5.4 Analysis II: Distribution Alignment and Representation Diagnostics

We further examine whether the benchmark gains are accompanied by reduced modality-level mismatch and preserved representation structure. Improved downstream performance does not necessarily imply that image and text representations become more closely aligned at the distribution level, or that their underlying structure is maintained. To investigate these aspects, we visualize spectral magnitude profiles of empirical characteristic functions to assess cross-modal distributional discrepancies, alongside shared-space t-SNE [79] embeddings to qualitatively examine modality separation. Both analyses are conducted on the COCO validation split.

Figure 5 shows that SpecAlign reduces the spectral magnitude gap between image and text embeddings across the evaluated frequencies. The observed reduction indicates closer agreement in the magnitude profiles of their empirical characteristic functions. However, magnitude agreement alone does not establish phase agreement or equality of the full characteristic functions, and therefore should not be interpreted as complete distributional alignment. Figure 6 provides a complementary visualization of the shared embedding space, showing reduced modality separation after alignment. While this offers qualitative evidence of improved cross-modal alignment, the visualization alone does not establish whether the underlying semantic structure is preserved.

**Representation Diversity.** To examine potential representation degeneration, we compute the effective rank of centered, sample-wise normalized feature matrices: $\mathrm{erank}(Z) = \exp(-\sum_j p_j \log p_j)$, where $p_j = \sigma_j / \sum_l \sigma_l$ and $\sigma_j$ are the singular values of $Z$. As shown in Table 5, the image effective rank decreases from 193.97 to 158.10, while the text effective rank increases from 111.43 to 148.24. Both remain substantially above one, providing evidence against one-dimensional collapse.

**Distribution Alignment and Semantic Discrimination.** We further evaluate distribution alignment and semantic discrimination on ImageNet validation. As shown in Table 6, SpecAlign achieves a lower sliced Wasserstein distance (SWD) and a higher Fisher ratio than STRUCTURE. SWD mea-

sures cross-modal distribution discrepancy, while the Fisher ratio measures inter-class separation relative to intra-class variation. These results suggest that SpecAlign reduces modality-level mismatch while retaining class-discriminative structure.

Table 5: Effective rank of image and text representations before and after alignment.

| Modality | Before | After |
|---|---|---|
| Image | 193.97 | 158.10 |
| Text | 111.43 | 148.24 |

Table 6: Distribution alignment and semantic discrimination on ImageNet validation.

| Metric | STRUCTURE | SpecAlign |
|---|---|---|
| SWD ↓ | 0.011475 | **0.010525** |
| Fisher ratio ↑ | 1.28172 | **1.36885** |

Together with the paired contrastive objective, these diagnostics support meaningful representation alignment rather than trivial collapse, although they do not rule out every possible form of representation degeneration. Additional data-scale and sensitivity analyses are provided in Appendix B.

### 5.5 Ablation Study

We ablate SpecAlign's spectral probing design in Table 7, varying whether frequency directions are learned and how radii are parameterized. Full SpecAlign performs best on both average zero-shot classification and Flickr30K image-to-text retrieval, showing that the gains do not come from random probes alone. Replacing the shared radius with sample-dependent radii hurts both metrics, while changing the radius rule under fixed random directions has only minor effect, indicating that learned directions are the main source of improvement. The retrieval column complements classification by testing ranking quality rather than prototype calibration. The consistent ordering across both metrics suggests that SDRS improves not only class calibration but also cross-modal neighborhood structure. Overall, the ablation supports the intended SDRS design: directions should be adaptive, while the radius should remain shared to keep spectral comparisons at a controlled scale.

Table 7: Ablation study of SpecAlign. We report average zero-shot classification accuracy and Flickr30K image-to-text Recall@1.

| **Variant** | **Learned freq.** | **Radius** | **ZS Avg.** | **Flickr30K I2T@1** |
|---|---|---|---|---|
| Full SpecAlign | ✓ | Shared | **43.5** | **67.4** |
| Fixed random freq. | ✗ | Raw | 41.8 | 65.2 |
| Sample-dependent radius | ✓ | Sample-dependent | 40.8 | 64.0 |
| Fixed random + shared radius | ✗ | Shared | 41.7 | 65.6 |
| Fixed random + sample-dependent radius | ✗ | Sample-dependent | 41.9 | 65.3 |

## 6 Conclusion

We proposed SpecAlign, a spectral characteristic-function matching framework for low-resource multimodal alignment, with distribution-level matching as its primary supervision mechanism. SpecAlign supervises modality-level distribution mismatch by using SDRS-guided probes to match image and text distributions in frequency space, while keeping pretrained encoders frozen. Across classification, retrieval, distribution-shift, and diagnostic evaluations, controlled spectral probing remains effective when paired data are scarce in the evaluated benchmark settings. The analysis further shows that probe stability matters: separating direction search from radius control gives the spectral objective a more consistent comparison scale under limited mini-batch estimates. A limitation is that our evaluation focuses on standard vision-language benchmarks, which may not fully reflect fine-grained shifts, long-tail categories, or noisy web-scale pairs. Future work will extend SpecAlign to broader domain-shift and noisy-pair settings with heterogeneous data quality and modality imbalance, and evaluate its behavior under these conditions.

## Acknowledgements

This paper was supported by the National Natural Science Foundation of China (No. 62376026, 62388101) and Beijing Nova Program (No. 20230484296).

## References


[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

[3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024.

[4] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2013.

[5] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018.

[6] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023.

[8] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

[9] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[10] Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *ICLR*, 2024.

[11] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *EMNLP*, 2022.

[12] Xiaoyang Wei, Camille Kurtz, and Florence Cloppet. Relaxing binary constraints in contrastive vision-language medical representation learning. In *WACV*, 2025.

[13] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022.

[14] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *CVPR*, 2023.

[15] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. PromptKD: Unsupervised prompt distillation for vision-language models. In *CVPR*, 2024.

[16] Jinyoung Park, Juyeon Ko, and Hyunwoo J. Kim. Prompt learning via meta-regularization. In *CVPR*, 2024.

[17] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *ECCV*, 2022.

[18] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, 2024.

[19] Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. MMA: Multi-modal adapter for vision-language models. In *CVPR*, 2024.

[20] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *ICML*, 2019.

[21] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

[22] Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Ankit Singh, and Noel E O'Connor. Harnessing frozen unimodal encoders for flexible multimodal alignment. In *CVPR*, 2025.

[23] Fabian Gröger, Shuo Wen, Huyen Le, and Maria Brbić. With limited data for multimodal alignment, let the STRUCTURE guide you. In *NeurIPS*, 2025.

[24] Zaid Khan and Yun Fu. Contrastive alignment of vision to language through parameter-efficient transfer learning. *ICLR*, 2023.

[25] Po-han Li, Sandeep P Chinchali, and Ufuk Topcu. Csa: Data-efficient mapping of unimodal features to multimodal features. In *ICLR*, 2025.

[26] Eugene Lukacs. *Characteristic Functions*. Hafner Publishing Company, 2 edition, 1970. ISBN 9780852641705.

[27] Jun Yu. Empirical characteristic function estimation and its applications. *Econometric reviews*, 2004.

[28] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *NeurIPS*, 2020.

[29] Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In *NeurIPS*, 2016.

[30] Shaobo Wang, Yicun Yang, Zhiyuan Liu, Chenghao Sun, Xuming Hu, Conghui He, and Linfeng Zhang. Dataset distillation with neural characteristic function: A minmax perspective. In *CVPR*, 2025.

[31] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022.

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[33] Hongyang Lei, Xiaolong Cheng, Qi Qin, Dan Wang, Kun Fan, Huazhen Huang, Qingqing Gu, Yetao Wu, Zhonglin Jiang, Yong Chen, and Luo Ji. M3-JEPA: Multimodal alignment via multi-gate MoE based on the joint-embedding predictive architecture. In *ICML*, 2025.

[34] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022.

[35] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*, 2023.

[36] Jameel Hassan Abdul Samadh, Hanan Gani, Noor Hazim Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Khan, and Salman Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *NeurIPS*, 2023.

[37] Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. In *ICLR*, 2024.

[38] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, 2023.

[39] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *CVPR*, 2023.

[40] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *ICML*, 2023.

[41] Avik Pal, Max Van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *ICLR*, 2025.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[43] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[44] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[45] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

[46] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024.

[47] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[48] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[49] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. In *CVPR*, 2023.

[50] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023. URL `https://arxiv.org/abs/2303.15389`.

[51] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022.

[52] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[53] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. OneLLM: One framework to align all modalities with language. In *CVPR*, 2024.

[54] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. LanguageBind: Extending video-language pretraining to N-modality by language-based semantic alignment. In *ICLR*, 2024.

[55] Jiaang Li, Yova Kementchedjhieva, Constanza Fierro, and Anders Søgaard. Do vision and language models share concepts? a vector space alignment study. *TACL*, 2024.

[56] Mustafa Shukor and Matthieu Cord. Implicit multimodal alignment: On the generalization of frozen LLMs to multimodal inputs. In *NeurIPS*, 2024.

[57] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

[58] Wei Chen, Xingyu Guo, Shuang Li, Zhao Zhang, Yan Zhong, Fuzhen Zhuang, et al. Learning adaptive distribution alignment with neural characteristic function for graph domain adaptation. *arXiv preprint arXiv:2602.10489*, 2026.

[59] Chuanhao Sun, Zhihang Yuan, Kai Xu, Luo Mai, N. Siddharth, Shuo Chen, and Mahesh K. Marina. Learning high-frequency functions made easy with sinusoidal positional encoding. In *ICML*, 2024.

[60] Talay M Cheema and Carl Edward Rasmussen. Integrated variational fourier features for fast spatial modelling with gaussian processes. *TMLR*, 2024.

[61] Jinyue Tian, Hui Xue, Yanfang Xue, and Pengfei Fang. Copula-nested spectral kernel network. In *ICML*, 2024.

[62] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2011.

[63] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[64] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

[65] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *IJCNN*, 2011.

[66] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.

[67] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014.

[68] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshop*, 2004.

[69] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[70] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[71] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[72] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

[73] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021.

[74] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021.

[75] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.

[76] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, 2019.

[77] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014.

[78] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[79] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008.

# A Additional Theoretical Details

This appendix collects proof details and derivations for the main theoretical statements in Section 4. The goal is to support the mechanism-level analysis of SpecAlign: SDRS provides controlled frequency probing, while the spectral objective measures cross-modal discrepancy through the frequencies it probes.

**Proof sketch of Lemma 4.2.** Assume $r_0 > 0$ and $\mathbf{u}_m \neq \mathbf{0}$ for each sampled direction; in implementation, the normalization can be stabilized by adding a small numerical constant. By construction, each sampled frequency has the form

$$\boldsymbol{\tau}_m = r\hat{\mathbf{u}}_m, \qquad \hat{\mathbf{u}}_m = \frac{\mathbf{u}_m}{\|\mathbf{u}_m\|_2}. \tag{19}$$

Hence $\|\hat{\mathbf{u}}_m\|_2 = 1$, and because the clipped radius is positive, we have

$$\|\boldsymbol{\tau}_m\|_2 = \|r\hat{\mathbf{u}}_m\|_2 = r. \tag{20}$$

Therefore all sampled frequencies in the same forward pass share exactly the same norm, so their norm variance is zero. The scale bound

$$r_0\, r_{\min} \leq \|\boldsymbol{\tau}_m\|_2 \leq r_0\, r_{\max} \tag{21}$$

follows directly from the clipped-radius definition

$$r = r_0 \cdot \mathrm{clip}(\exp(\alpha), r_{\min}, r_{\max})\,. \tag{22}$$

This establishes that the shared-radius parameterization eliminates norm variation within a forward pass while keeping the overall probing scale under explicit control.

Next, let $\hat{\mathbf{u}} = \mathbf{u}/\|\mathbf{u}\|_2$ for $\mathbf{u} \neq \mathbf{0}$. The standard differentiation rule for normalized vectors gives

$$\frac{\partial \hat{\mathbf{u}}}{\partial \mathbf{u}} = \frac{1}{\|\mathbf{u}\|_2}\left(I - \hat{\mathbf{u}}\hat{\mathbf{u}}^\top\right). \tag{23}$$

Since $\boldsymbol{\tau} = r\hat{\mathbf{u}}$ and $r$ is independent of the sample-specific direction variable $\mathbf{u}$ inside this local derivative, we obtain

$$\frac{\partial \boldsymbol{\tau}}{\partial \mathbf{u}} = \frac{r}{\|\mathbf{u}\|_2}\left(I - \hat{\mathbf{u}}\hat{\mathbf{u}}^\top\right). \tag{24}$$

Now consider a purely radial perturbation $\delta\mathbf{u} = c\,\hat{\mathbf{u}}$. Then

$$\left(I - \hat{\mathbf{u}}\hat{\mathbf{u}}^\top\right)\delta\mathbf{u} = c\left(I - \hat{\mathbf{u}}\hat{\mathbf{u}}^\top\right)\hat{\mathbf{u}} = c\left(\hat{\mathbf{u}} - \hat{\mathbf{u}}\,\hat{\mathbf{u}}^\top\hat{\mathbf{u}}\right) = \mathbf{0}, \tag{25}$$

because $\hat{\mathbf{u}}^\top\hat{\mathbf{u}} = 1$. Therefore

$$\frac{\partial \boldsymbol{\tau}}{\partial \mathbf{u}}\,\delta\mathbf{u} = \mathbf{0}. \tag{26}$$

Thus the normalization step removes first-order sensitivity to purely radial perturbations and preserves only tangent-space motion along the hypersphere.

Finally, if the aligned embeddings are cosine-normalized, then $\|\mathbf{z}\|_2 = 1$. Therefore

$$|\boldsymbol{\tau}^\top\mathbf{z}| \leq \|\boldsymbol{\tau}\|_2\|\mathbf{z}\|_2 = \|\boldsymbol{\tau}\|_2 \leq r_0\, r_{\max}, \tag{27}$$

which proves the characteristic-function input bound and completes the proof sketch.

**Proof sketch of Proposition 3.1.** Recall that $P_v$ and $P_t$ denote the image and text embedding distributions, with characteristic functions $\phi_v$ and $\phi_t$, respectively. If $\mathcal{C}_\psi(P_v, P_t) = 0$, then by the definition of the spectral discrepancy,

$$|\phi_v(\boldsymbol{\tau}) - \phi_t(\boldsymbol{\tau})|^2 = 0 \qquad \text{for } F_\psi\text{-a.e. } \boldsymbol{\tau}. \tag{28}$$

Let

$$g(\boldsymbol{\tau}) = \phi_v(\boldsymbol{\tau}) - \phi_t(\boldsymbol{\tau}). \tag{29}$$

Since characteristic functions are continuous, $g$ is continuous on $\mathbb{R}^k$. We first show that $g$ vanishes on the open set $\Omega$. Suppose otherwise. Then there exists $\boldsymbol{\tau}_0 \in \Omega$ such that $g(\boldsymbol{\tau}_0) \neq 0$. By continuity,

there exists an open neighborhood $U \subset \Omega$ of $\boldsymbol{\tau}_0$ on which $|g|$ is bounded away from zero. Because $\Omega \subset \operatorname{supp}(F_\psi)$, this non-empty open set satisfies $F_\psi(U) > 0$, which contradicts the fact that $|g|^2 = 0$ for $F_\psi$-a.e. $\boldsymbol{\tau}$. Therefore,

$$\phi_v(\boldsymbol{\tau}) = \phi_t(\boldsymbol{\tau}) \qquad \text{for all } \boldsymbol{\tau} \in \Omega. \tag{30}$$

By the local moment-generating-function assumption, both characteristic functions are real analytic. Since their difference vanishes on a non-empty open subset $\Omega \subset \mathbb{R}^k$, the identity theorem for real-analytic functions implies

$$\phi_v(\boldsymbol{\tau}) = \phi_t(\boldsymbol{\tau}) \qquad \text{for all } \boldsymbol{\tau} \in \mathbb{R}^k. \tag{31}$$

The uniqueness theorem of characteristic functions then yields

$$P_v = P_t. \tag{32}$$

This proves the proposition.

**Additional discussion for Proposition 3.1.** This proposition is an idealized rich-probing statement, not a finite-sample recovery guarantee. It only states that, under rich probing and the stated analyticity condition, characteristic-function matching is a distribution-identifying criterion. In the main method, SDRS deliberately restricts probing to a controlled spectral scale, so the practical role of SpecAlign is to reduce the mismatch exposed by stable learned probes rather than to guarantee full distribution recovery from finitely many frequencies.

**Proof sketch of Proposition 4.1.** Recall from Eq. (8) that

$$d_{\mathrm{CF}}(\boldsymbol{\tau}) = A(\boldsymbol{\tau}) + B(\boldsymbol{\tau}), \tag{33}$$

where

$$A(\boldsymbol{\tau}) = \big(\rho_v(\boldsymbol{\tau}) - \rho_t(\boldsymbol{\tau})\big)^2, \qquad B(\boldsymbol{\tau}) = 2\rho_v(\boldsymbol{\tau})\rho_t(\boldsymbol{\tau})\big(1 - \cos(a_v(\boldsymbol{\tau}) - a_t(\boldsymbol{\tau}))\big). \tag{34}$$

We derive the low-frequency behavior by expanding the amplitude and phase parts separately.

**Step 1: Local expansion of the characteristic function.** Assume that the relevant third moments are finite. For each modality $u \in \{v, t\}$, the cumulant expansion around $\boldsymbol{\tau} = \mathbf{0}$ gives

$$\log \phi_u(\boldsymbol{\tau}) = \mathrm{i}\, \boldsymbol{\mu}_u^\top \boldsymbol{\tau} - \frac{1}{2}\boldsymbol{\tau}^\top \Sigma_u \boldsymbol{\tau} + O(\|\boldsymbol{\tau}\|^3). \tag{35}$$

For $\|\boldsymbol{\tau}\|$ sufficiently small, $\phi_u(\boldsymbol{\tau}) \neq 0$, so the amplitude-phase decomposition is well defined. Taking imaginary and real parts yields

$$a_u(\boldsymbol{\tau}) = \boldsymbol{\mu}_u^\top \boldsymbol{\tau} + O(\|\boldsymbol{\tau}\|^3), \qquad \log \rho_u(\boldsymbol{\tau}) = -\frac{1}{2}\boldsymbol{\tau}^\top \Sigma_u \boldsymbol{\tau} + O(\|\boldsymbol{\tau}\|^3). \tag{36}$$

Exponentiating the second relation gives

$$\rho_u(\boldsymbol{\tau}) = 1 - \frac{1}{2}\boldsymbol{\tau}^\top \Sigma_u \boldsymbol{\tau} + O(\|\boldsymbol{\tau}\|^3). \tag{37}$$

Therefore, with

$$\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_v - \boldsymbol{\mu}_t, \qquad \Delta\Sigma = \Sigma_v - \Sigma_t, \tag{38}$$

we obtain

$$a_v(\boldsymbol{\tau}) - a_t(\boldsymbol{\tau}) = \boldsymbol{\tau}^\top \Delta\boldsymbol{\mu} + O(\|\boldsymbol{\tau}\|^3), \tag{39}$$

and

$$\rho_v(\boldsymbol{\tau}) - \rho_t(\boldsymbol{\tau}) = -\frac{1}{2}\boldsymbol{\tau}^\top \Delta\Sigma\, \boldsymbol{\tau} + O(\|\boldsymbol{\tau}\|^3). \tag{40}$$

**Step 2: Expansion of the amplitude term.** Using the previous display,

$$A(\boldsymbol{\tau}) = \left(-\frac{1}{2}\boldsymbol{\tau}^\top \Delta\Sigma\, \boldsymbol{\tau} + O(\|\boldsymbol{\tau}\|^3)\right)^2. \tag{41}$$

Since $\boldsymbol{\tau}^\top \Delta\Sigma\, \boldsymbol{\tau} = O(\|\boldsymbol{\tau}\|^2)$, we have

$$A(\boldsymbol{\tau}) = \frac{1}{4}\big(\boldsymbol{\tau}^\top \Delta\Sigma\, \boldsymbol{\tau}\big)^2 + O(\|\boldsymbol{\tau}\|^5). \tag{42}$$

Thus the amplitude component exposes covariance mismatch at fourth order.

**Step 3: Expansion of the phase term.** Let

$$\delta a(\boldsymbol{\tau}) = a_v(\boldsymbol{\tau}) - a_t(\boldsymbol{\tau}). \tag{43}$$

From the expansion above,

$$\delta a(\boldsymbol{\tau}) = \boldsymbol{\tau}^\top \Delta\boldsymbol{\mu} + O(\|\boldsymbol{\tau}\|^3), \qquad \rho_v(\boldsymbol{\tau})\rho_t(\boldsymbol{\tau}) = 1 + O(\|\boldsymbol{\tau}\|^2). \tag{44}$$

Using

$$1 - \cos x = \frac{1}{2}x^2 + O(x^4), \qquad x \to 0, \tag{45}$$

we get

$$2\big(1 - \cos \delta a(\boldsymbol{\tau})\big) = \delta a(\boldsymbol{\tau})^2 + O(\|\boldsymbol{\tau}\|^4). \tag{46}$$

Since $\delta a(\boldsymbol{\tau}) = O(\|\boldsymbol{\tau}\|)$,

$$\delta a(\boldsymbol{\tau})^2 = \big(\boldsymbol{\tau}^\top \Delta\boldsymbol{\mu}\big)^2 + O(\|\boldsymbol{\tau}\|^4). \tag{47}$$

Therefore

$$B(\boldsymbol{\tau}) = \big(1 + O(\|\boldsymbol{\tau}\|^2)\big)\Big(\big(\boldsymbol{\tau}^\top \Delta\boldsymbol{\mu}\big)^2 + O(\|\boldsymbol{\tau}\|^4)\Big) = \big(\boldsymbol{\tau}^\top \Delta\boldsymbol{\mu}\big)^2 + O(\|\boldsymbol{\tau}\|^4). \tag{48}$$

Thus the phase component first reveals the mean gap at second order.

**Step 4: Combine the two parts.** Combining the amplitude and phase expansions gives

$$d_{\mathrm{CF}}(\boldsymbol{\tau}) = \big(\boldsymbol{\tau}^\top \Delta\boldsymbol{\mu}\big)^2 + O(\|\boldsymbol{\tau}\|^4). \tag{49}$$

More specifically, the decomposition shows that

$$A(\boldsymbol{\tau}) = \frac{1}{4}\big(\boldsymbol{\tau}^\top \Delta\Sigma\, \boldsymbol{\tau}\big)^2 + O(\|\boldsymbol{\tau}\|^5), \qquad B(\boldsymbol{\tau}) = \big(\boldsymbol{\tau}^\top \Delta\boldsymbol{\mu}\big)^2 + O(\|\boldsymbol{\tau}\|^4). \tag{50}$$

Therefore low frequencies first reveal the phase-driven mean gap, while the amplitude branch exposes covariance mismatch at fourth order. In the special case where $\Delta\boldsymbol{\mu} = \mathbf{0}$, the leading phase contribution vanishes and

$$d_{\mathrm{CF}}(\boldsymbol{\tau}) = \frac{1}{4}\big(\boldsymbol{\tau}^\top \Delta\Sigma\, \boldsymbol{\tau}\big)^2 + O(\|\boldsymbol{\tau}\|^5), \tag{51}$$

up to higher-order phase terms. This formalizes the claim that low-frequency characteristic-function matching captures coarse distributional discrepancies, with mean offsets appearing through phase and covariance mismatch appearing through amplitude.

**Translation example for Proposition 4.1.** Suppose the text-side embedding distribution is a translation of the image-side embedding distribution, i.e.,

$$\mathbf{z}_t \overset{d}{=} \mathbf{z}_v + \mathbf{c} \tag{52}$$

for some nonzero shift $\mathbf{c}$. Then

$$\phi_t(\boldsymbol{\tau}) = \mathbb{E}\left[e^{\mathrm{i}\boldsymbol{\tau}^\top(\mathbf{z}_v + \mathbf{c})}\right] = e^{\mathrm{i}\boldsymbol{\tau}^\top \mathbf{c}}\phi_v(\boldsymbol{\tau}). \tag{53}$$

Therefore the amplitudes agree,

$$|\phi_t(\boldsymbol{\tau})| = |\phi_v(\boldsymbol{\tau})|, \tag{54}$$

while the phase is shifted by $\boldsymbol{\tau}^\top \mathbf{c}$. This illustrates why phase is sensitive to global cross-modal offsets, whereas amplitude is unchanged under a pure translation.

**Formal assumptions for Theorem 4.3.** The theorem is stated for a full-dimensional probing distribution with density $p_\psi(\boldsymbol{\tau})$ with respect to Lebesgue measure. This should be read as a coverage-based stability statement. The exact shared-radius SDRS used in one forward pass probes a controlled spectral shell, while a smoothed or radially jittered probing distribution satisfies the density assumption directly.

Let $h : \mathbb{R}^k \to \mathbb{R}$ be a scoring function with Fourier transform $\widehat{h}$ such that the Fourier inversion formula holds for $h, \widehat{h} \in L^1(\mathbb{R}^k)$, and

$$\int_{\mathbb{R}^k} \frac{|\widehat{h}(\boldsymbol{\tau})|^2}{p_\psi(\boldsymbol{\tau})}\, d\boldsymbol{\tau} < \infty. \tag{55}$$

We use the convention that $|\widehat{h}(\boldsymbol{\tau})|^2/p_\psi(\boldsymbol{\tau}) = +\infty$ whenever $p_\psi(\boldsymbol{\tau}) = 0$ and $\widehat{h}(\boldsymbol{\tau}) \neq 0$. Thus the condition requires the probing density to cover the task-relevant Fourier support of $h$.

**Proof sketch of Theorem 4.3.** By Fourier inversion, we may write

$$h(\mathbf{z}) = \int_{\mathbb{R}^k} \widehat{h}(\boldsymbol{\tau}) e^{\mathrm{i}\boldsymbol{\tau}^\top \mathbf{z}}\, d\boldsymbol{\tau}, \tag{56}$$

where the normalization constant is absorbed into the definition of $\widehat{h}$. Since $\widehat{h} \in L^1(\mathbb{R}^k)$ and $|\phi_u(\boldsymbol{\tau})| \le 1$, Fubini's theorem allows us to exchange expectation and integration, yielding

$$\mathbb{E}_{P_v}[h(\mathbf{z})] = \int_{\mathbb{R}^k} \widehat{h}(\boldsymbol{\tau})\phi_v(\boldsymbol{\tau})\, d\boldsymbol{\tau}, \qquad \mathbb{E}_{P_t}[h(\mathbf{z})] = \int_{\mathbb{R}^k} \widehat{h}(\boldsymbol{\tau})\phi_t(\boldsymbol{\tau})\, d\boldsymbol{\tau}. \tag{57}$$

Subtracting the two identities gives

$$\mathbb{E}_{P_v}[h(\mathbf{z})] - \mathbb{E}_{P_t}[h(\mathbf{z})] = \int_{\mathbb{R}^k} \widehat{h}(\boldsymbol{\tau})\big(\phi_v(\boldsymbol{\tau}) - \phi_t(\boldsymbol{\tau})\big)\, d\boldsymbol{\tau}. \tag{58}$$

Insert $p_\psi(\boldsymbol{\tau})^{-1/2} p_\psi(\boldsymbol{\tau})^{1/2}$ inside the integral:

$$\mathbb{E}_{P_v}[h(\mathbf{z})] - \mathbb{E}_{P_t}[h(\mathbf{z})] = \int_{\mathbb{R}^k} \frac{\widehat{h}(\boldsymbol{\tau})}{\sqrt{p_\psi(\boldsymbol{\tau})}}\big(\phi_v(\boldsymbol{\tau}) - \phi_t(\boldsymbol{\tau})\big)\sqrt{p_\psi(\boldsymbol{\tau})}\, d\boldsymbol{\tau}. \tag{59}$$

Applying Cauchy–Schwarz gives

$$|\mathbb{E}_{P_v}[h(\mathbf{z})] - \mathbb{E}_{P_t}[h(\mathbf{z})]| \le \left(\int_{\mathbb{R}^k} \frac{|\widehat{h}(\boldsymbol{\tau})|^2}{p_\psi(\boldsymbol{\tau})}\, d\boldsymbol{\tau}\right)^{1/2} \left(\int_{\mathbb{R}^k} |\phi_v(\boldsymbol{\tau}) - \phi_t(\boldsymbol{\tau})|^2 p_\psi(\boldsymbol{\tau})\, d\boldsymbol{\tau}\right)^{1/2}. \tag{60}$$

The second factor is exactly $\sqrt{\mathcal{C}_\psi(P_v, P_t)}$, so

$$|\mathbb{E}_{P_v}[h(\mathbf{z})] - \mathbb{E}_{P_t}[h(\mathbf{z})]| \le \left(\int_{\mathbb{R}^k} \frac{|\widehat{h}(\boldsymbol{\tau})|^2}{p_\psi(\boldsymbol{\tau})}\, d\boldsymbol{\tau}\right)^{1/2} \sqrt{\mathcal{C}_\psi(P_v, P_t)}. \tag{61}$$

This proves the theorem. In particular, as long as the Fourier energy of $h$ is not concentrated in regions where $p_\psi(\boldsymbol{\tau})$ is extremely small, reducing $\mathcal{C}_\psi(P_v, P_t)$ bounds the downstream cross-modal expectation gap under the stated coverage condition. Thus the result should be read as a conditional stability statement for task-relevant shared-space functions rather than as an unconditional generalization guarantee.

**Scope of Theorem 4.3.** The bound is informative only when the probing density covers the Fourier support relevant to the downstream score. If $\widehat{h}$ concentrates in regions where $p_\psi$ is very small, the prefactor can become large even when $\mathcal{C}_\psi(P_v, P_t)$ is small. This is why the theorem is best interpreted as a coverage-based explanation of score stability, not as a universal guarantee for arbitrary downstream functions.

**Proof sketch of Corollary 4.4.** For each class $y$, let $P_v^y$ and $P_t^y$ denote the class-conditional image and text embedding distributions. Assume, as stated in the corollary, that for every class score $s_c$ one has the class-conditional deviation bound

$$\big|\mathbb{E}_{P_v^y}[s_c(\mathbf{z})] - \mathbb{E}_{P_t^y}[s_c(\mathbf{z})]\big| \le B_c \sqrt{\mathcal{C}_\psi(P_v^y, P_t^y)}. \tag{62}$$

Let $B = \max_c B_c$. Then, for the correct class $y$,

$$\mathbb{E}_{P_t^y}[s_y(\mathbf{z})] \ge \mathbb{E}_{P_v^y}[s_y(\mathbf{z})] - B\sqrt{\mathcal{C}_\psi(P_v^y, P_t^y)}. \tag{63}$$

For any incorrect class $c \ne y$,

$$\mathbb{E}_{P_t^y}[s_c(\mathbf{z})] \le \mathbb{E}_{P_v^y}[s_c(\mathbf{z})] + B\sqrt{\mathcal{C}_\psi(P_v^y, P_t^y)}. \tag{64}$$

Taking the maximum over all incorrect classes yields

$$\max_{c \ne y} \mathbb{E}_{P_t^y}[s_c(\mathbf{z})] \le \max_{c \ne y} \mathbb{E}_{P_v^y}[s_c(\mathbf{z})] + B\sqrt{\mathcal{C}_\psi(P_v^y, P_t^y)}. \tag{65}$$

Subtracting this inequality from the lower bound for the correct class gives

$$\mathbb{E}_{P_t^y}[s_y(\mathbf{z})] - \max_{c\neq y} \mathbb{E}_{P_t^y}[s_c(\mathbf{z})] \geq \mathbb{E}_{P_v^y}[s_y(\mathbf{z})] - \max_{c\neq y} \mathbb{E}_{P_v^y}[s_c(\mathbf{z})] - 2B\sqrt{\mathcal{C}_\psi(P_v^y, P_t^y)}. \tag{66}$$

By the assumed visual-side expected margin,

$$\mathbb{E}_{P_v^y}[s_y(\mathbf{z})] - \max_{c\neq y} \mathbb{E}_{P_v^y}[s_c(\mathbf{z})] \geq \gamma_y, \tag{67}$$

we obtain

$$\mathbb{E}_{P_t^y}[s_y(\mathbf{z})] - \max_{c\neq y} \mathbb{E}_{P_t^y}[s_c(\mathbf{z})] \geq \gamma_y - 2B\sqrt{\mathcal{C}_\psi(P_v^y, P_t^y)}. \tag{68}$$

Therefore, if

$$2B\sqrt{\mathcal{C}_\psi(P_v^y, P_t^y)} < \gamma_y, \tag{69}$$

then

$$\mathbb{E}_{P_t^y}[s_y(\mathbf{z})] > \max_{c\neq y} \mathbb{E}_{P_t^y}[s_c(\mathbf{z})], \tag{70}$$

which proves that the expected class ordering is preserved for class $y$.

This corollary gives a conditional mechanism for robustness: when the score-drift condition holds class-conditionally, reducing the spectral discrepancy stabilizes expected shared-space class ordering across modalities. It does not imply a samplewise accuracy guarantee.

**Finite-sample discussion.** For a fixed frequency $\boldsymbol{\tau}$, the empirical characteristic function is an average of unit-modulus complex random variables:

$$\hat{\phi}_u(\boldsymbol{\tau}) = \frac{1}{B}\sum_{i=1}^{B} e^{\mathrm{i}\boldsymbol{\tau}^\top \mathbf{z}_u^i}. \tag{71}$$

Its real and imaginary parts are bounded, so standard concentration arguments imply

$$\left|\hat{\phi}_u(\boldsymbol{\tau}) - \phi_u(\boldsymbol{\tau})\right| = O_p(B^{-1/2}) \tag{72}$$

for a fixed frequency probe. For a finite set of $M$ non-adaptive probes, a standard union-bound argument gives

$$\max_{1\leq m\leq M} \left|\hat{\phi}_u(\boldsymbol{\tau}_m) - \phi_u(\boldsymbol{\tau}_m)\right| = O_p\left(\sqrt{\frac{\log M}{B}}\right). \tag{73}$$

Thus empirical characteristic-function estimates provide stable summaries of the population-level discrepancy for fixed or finitely many non-adaptive probes. For learned adaptive probes, this discussion should be interpreted as an intuition for the statistical stability of the empirical estimates rather than as a uniform finite-sample guarantee over the entire learned probe family.

# B Additional Experimental Analysis

## B.1 Robustness Across Training Data Scales

We test whether SpecAlign remains effective as paired supervision decreases. Figure 7 reports results on six representative benchmarks, using progressively reduced training pairs. SpecAlign outperforms the baseline on the evaluated data scales and datasets, indicating that its advantage is not tied to the full-data regime.

## B.2 Sensitivity to Magnitude and Phase Weighting

We examine the implementation-level weighting between the magnitude and phase components of the spectral objective in Figure 8. Performance is highest in this sweep when the two terms are kept in a relatively balanced regime, whereas over-emphasizing either component leads to a noticeable drop in classification accuracy. This trend is consistent with the amplitude-phase interpretation in Section 4.1: the magnitude component primarily constrains spectral energy mismatch, while the phase component is sensitive to cross-modal displacement and structural offset. Overweighting either component makes the objective capture only a partial view of the cross-modal gap, yielding a less effective alignment.

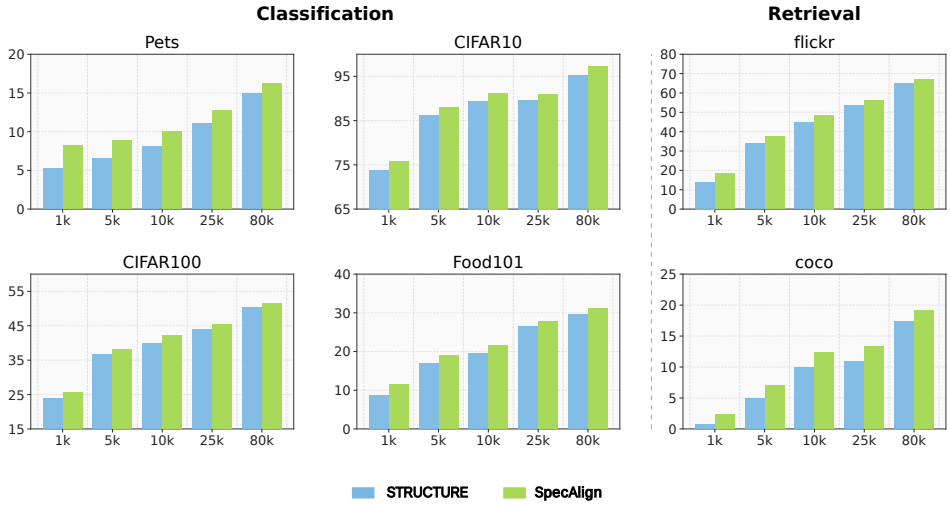



Figure 7: Performance trends under reduced training data scales across representative benchmarks.

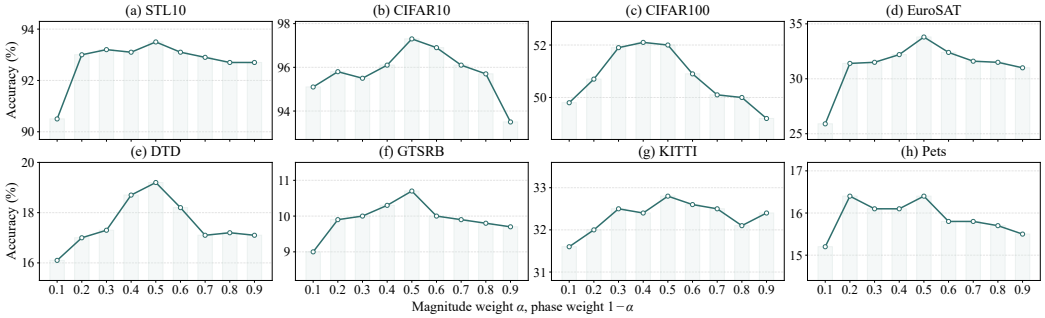



Figure 8: Sensitivity of classification accuracy to implementation-level magnitude/phase weighting. A balanced weighting strategy achieves the highest accuracy in this sweep, supporting the complementary roles of magnitude and phase information.

### B.3 Sensitivity to SDRS Sampler Batch Size

We further analyze the SDRS sampler batch size $t_{\text{batch}}$, which controls the number of frequency probes produced by the sampler in each forward pass and should not be confused with the paired training-data mini-batch size. As shown in Figure 9, the optimal sampler batch size is not identical across datasets, but performance remains stable over a wide range of values. Larger sampler batches remain competitive on most datasets, indicating that SpecAlign is not highly sensitive to precise tuning of the sampler-batch hyperparameter.

## C Implementation Details and Hyperparameter Settings

Unless otherwise stated, alignment training uses paired image-text examples from the MS-COCO 2014 training split and follows the data split and evaluation protocol of STRUCTURE; baseline numbers are reported from the shared protocol, while SpecAlign is trained and evaluated under the same frozen-encoder, paired-data, training-epoch, and evaluation settings.

### C.1 Dataset Details

Table 8 summarizes the datasets used in our evaluation suite. These benchmarks cover a broad range of visual recognition domains, including generic objects, fine-grained categories, food, textures, traffic signs, remote sensing images, human actions, and driving scenes. We also include ImageNet-

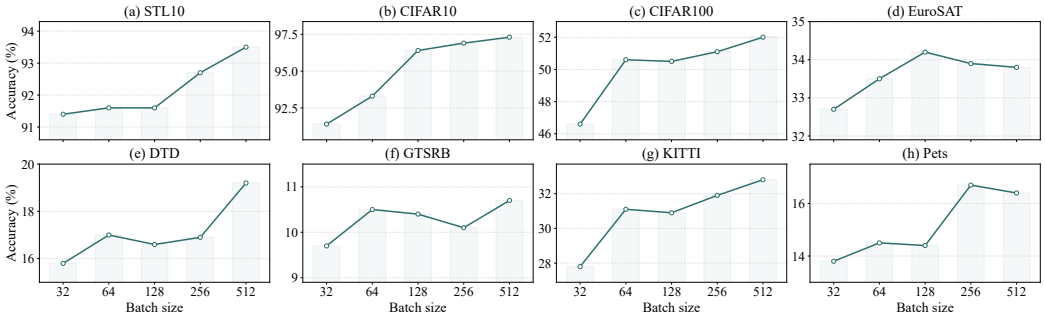



Figure 9: Sensitivity of zero-shot classification accuracy to the SDRS frequency-sampler batch size $t_{\text{batch}}$. This batch size controls the number of sampled frequency probes, not the paired training-data mini-batch size.

derived distribution-shift benchmarks to assess robustness under changes in visual appearance, and image–text retrieval datasets to evaluate cross-modal matching.

Table 8: Summary of datasets used in our evaluation. We report the task type, number of semantic categories when applicable, and the main visual or multimodal domain of each benchmark.

| Task type | Dataset | # Classes | Dataset description |
|---|---|---|---|
| Classification | STL10 [62] | 10 | Natural image classification with object-centric categories. |
| | CIFAR10 [63] | 10 | Small-scale natural image classification over common object classes. |
| | CIFAR100 [63] | 100 | Fine-grained natural image classification with 100 object categories. |
| | EuroSAT [64] | 10 | Land-use and land-cover classification from satellite images. |
| | GTSRB [65] | 43 | Traffic sign classification under real-world imaging conditions. |
| | Oxford-IIIT Pets [66] | 37 | Fine-grained classification of cat and dog breeds. |
| | Food101 [67] | 101 | Fine-grained food image classification. |
| | Caltech101 [68] | 101 | Object category recognition across diverse visual concepts. |
| | ImageNet [69] | 1000 | Large-scale object recognition benchmark. |
| | UCF101 [70] | 101 | Human action recognition dataset constructed from video data. |
| | KITTI Distance [71] | 4 | Four-category distance classification using KITTI driving imagery. |
| | DTD [72] | 47 | Texture recognition benchmark with describable texture categories. |
| Robustness | ImageNet-A [73] | 200 | ImageNet-compatible natural adversarial examples. |
| | ImageNet-R [74] | 200 | ImageNet-compatible artistic renditions and non-photorealistic images. |
| | ImageNet-V2 [76] | 1000 | Reconstructed ImageNet test distribution for measuring dataset shift. |
| | ImageNet-Sketch [75] | 1000 | Sketch-style images following the ImageNet label space. |
| Retrieval | Flickr30K [77] | N/A | Image–caption dataset for bidirectional image–text retrieval. |
| | MS-COCO [78] | N/A | Large-scale image–caption dataset with multiple captions per image. |

### C.2 Backbone Encoders and Alignment Heads

We use a frozen DINOv2 vision encoder and a frozen RoBERTa text encoder to extract unimodal features. Following the comparison protocol in the main experiments, both modalities are mapped into the shared alignment space by the same gated low-rank alignment head, and the shared projection dimension is fixed to 512 throughout all experiments. We consider two layer-selection settings: a final-layer setting, where both encoders use their output-layer representations, and a Similar-layer setting, where both encoders use the most similar cross-modal layer pair from the baseline protocol. Following the original baseline setting, SpecAlign is applied under both settings. Concretely, the alignment head consists of a base linear projection together with a low-rank correction branch of the form down-projection $\rightarrow$ GELU $\rightarrow$ dropout $\rightarrow$ up-projection. A sample-wise sigmoid gate modulates the strength of this correction before it is added to the base projection, which allows the head to adapt the correction magnitude to each input while remaining lightweight. Unless otherwise stated, the two backbone encoders remain frozen, and only the alignment heads together with the spectral sampler are updated during training.

### C.3 Optimization Details

We optimize the trainable modules with AdamW. The learning rate of the alignment model is set to $1 \times 10^{-5}$, and all experiments are trained for 1000 epochs. Hyperparameters are selected on a validation split and then kept fixed across datasets and training scales unless otherwise specified.

### C.4 Training Procedure

Algorithm 1 summarizes one mini-batch training iteration of SpecAlign under the min-max objective in Section 4. The encoders remain frozen, while the alignment heads are updated to minimize the total objective and the SDRS sampler is updated to maximize the spectral discrepancy it exposes.

**Algorithm 1** One training iteration of SpecAlign with SDRS

**Require:** Paired mini-batch $\{(\mathbf{x}_v^i, \mathbf{x}_t^i)\}_{i=1}^B$, frozen encoders $(f_v, f_t)$
**Require:** Alignment heads $(g_v, g_t)$ with parameters $(\theta_v, \theta_t)$, SDRS sampler $s_\psi$
**Require:** Number of probes $M$, temperature $\kappa$, weights $\alpha_{\mathrm{cf}}, \beta_{\mathrm{cf}}, \lambda_{\mathrm{SpecAlign}}$
$\bar{\mathbf{z}}_q^i \leftarrow n(g_q(f_q(\mathbf{x}_q^i)))$ for $q \in \{v, t\}$ and $i = 1, \ldots, B$ by Eq. (1)
Compute $\mathcal{L}_{\mathrm{CLIP}}$ by Eq. (12)
Sample $\boldsymbol{\epsilon}_m \sim \mathcal{N}(0, I)$ for $m = 1, \ldots, M$; set $r \leftarrow r_0 \cdot \mathrm{clip}(\exp(\alpha), r_{\min}, r_{\max})$
**for** $m = 1, \ldots, M$ **do**
    $\mathbf{u}_m \leftarrow W_o h_\psi(\boldsymbol{\epsilon}_m) + b_o$, $\boldsymbol{\tau}_m \leftarrow r\, \mathbf{u}_m / \|\mathbf{u}_m\|_2$
    $\hat{\phi}_q(\boldsymbol{\tau}_m) \leftarrow \frac{1}{B} \sum_{i=1}^B e^{\mathrm{i}\, \boldsymbol{\tau}_m^\top \bar{\mathbf{z}}_q^i}$ for $q \in \{v, t\}$
    Compute $\mathcal{L}_{\mathrm{amp}}^{(m)}$ and $\mathcal{L}_{\mathrm{pha}}^{(m)}$ by Eqs. (10) and (11)
**end for**
Compute $\mathcal{L}_{\mathrm{SpecAlign}}$ by Eq. (9); set $\mathcal{L}_{\mathrm{total}} \leftarrow \mathcal{L}_{\mathrm{CLIP}} + \lambda_{\mathrm{SpecAlign}} \mathcal{L}_{\mathrm{SpecAlign}}$
Update $(\theta_v, \theta_t)$ by descent on $\mathcal{L}_{\mathrm{total}}$ and $\psi$ by ascent on $\mathcal{L}_{\mathrm{SpecAlign}}$
**Ensure:** Updated alignment-head parameters $(\theta_v, \theta_t)$ and sampler parameters $\psi$

### C.5 Spectral Objective Hyperparameters

Table 9 summarizes the SpecAlign settings. We use equal spectral weights ($\alpha = \beta = 0.5$), 4096 sampled frequencies, a spectral loss scale and structure regularization coefficient of 1.0, and a 1-step warmup. The sampler uses a 1024-dimensional feature space, batch size 64, and $r_0 = 1.0$, with learning rate $10^{-3}$ and momentum 0.5.

Table 9: Implementation details and hyperparameter settings used in SpecAlign.

| Component | Setting |
|---|---|
| Shared projection dimension | 512 |
| Optimizer | AdamW |
| Learning rate | $1 \times 10^{-5}$ |
| Weight decay | $1 \times 10^{-4}$ |
| Training mini-batch size $B$ | 2048 |
| Contrastive temperature $\kappa$ | 0.05 |
| Loss weights $(\alpha, \beta)$ | $(0.5, 0.5)$ |
| Number of sampled frequencies $M$ | 4096 |
| Spectral loss scale | 1.0 |
| Structure regularization weight | 1.0 |
| Spectral warmup steps | 1 |
| Sampler feature dimension | 1024 |
| Sampler batch size | 64 |
| Base radius scale $r_0$ | 1.0 |
| Radius scale range $[r_{\min}, r_{\max}]$ | $[0.5, 2.0]$ |
| Sampler learning rate | $1 \times 10^{-3}$ |
| Sampler momentum | 0.5 |

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Section 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: See Section 6.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: See the Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed information on the experimental setting, including data splits, hyperparameters, and optimization details in Section 5 and Appendix C.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release the code and data upon publication.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide detailed information on the experimental setting details in the Experiment Section and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We did not report error bars due to computational constraints, but we will include them in future work.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide detailed information on the compute resources used for our experiments in 5 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research fully complies with the NeurIPS Code of Ethics. It uses only publicly available datasets and does not involve human subjects, private data, or sensitive content.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work does not have any foreseeable societal impact. It focuses on improving vision-language training methods without involving sensitive data, human subjects, or deployment-oriented applications. Therefore, we believe there are no direct or indirect risks associated with the proposed approach.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not involve the release of models or datasets that pose a high risk of misuse. It builds on publicly available datasets and does not include pretrained models or components with dual-use concerns.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper uses publicly available datasets that are properly cited with their original sources and licenses.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We will release the new assets introduced in the paper with detailed documentation, including information about training, license, and limitations.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: In this work, the LLM is utilized solely for writing and editing purposes. They do not influence the core methodology, scientific rigor, or the originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.