Stable Spectral Copula Alignment

For Robust Multimodal Learning

Authors: Hongkang Zhang, Shao-Lun Huang,
Yanlong Wang, Ercan Engin KURUOGLU

Affiliation: Tsinghua University

SSCA

Copula-stable dependence
+ spectral diagnostics

multimodal alignment can silently fail under deployment shift.

Core

An auditable copula-stable protocol with diagnostics and fallback.

Motivation & Background

Robust multimodal alignment under deployment shift

Why deployment shift breaks alignment

Current multimodal objectives often entangle cross-modal dependence with marginal-sensitive feature geometry.

Even when the semantic relationship across modalities remains informative, monotone or near-monotone feature distortions—such as gain changes, normalization drift, and moderate quantization—can silently corrupt alignment.

Raw-pipeline changes, including compression, tokenizer changes, and audio-codec shifts, further expose this fragility in realistic deployment.

Core gap Robustness is often improved empirically, but the connection between invariance assumptions, measurable diagnostics, and safe update / fallback decisions remains weak.

Key need: dependence that remains meaningful when marginals change.

Recent progress in this direction

2024

Robust VLM fine-tuning and calibration; parameter-efficient missing-modality adaptation(NeurIPS 2024; ICLR 2024)

2025

Dynamic modality adaptation for missingness, imbalance, and label-free VLM adaptation(Findings of EMNLP 2025)

2026

Reliability-aware multimodal TTA and VLA / MLLM robustness under multimodal perturbations (recent 2026 studies)

SSCA positioning from empirical robustness to a diagnostic dependence contract: update only when the contract is supported; otherwise fallback.

Dependence | Diagnostics | Calibration | Fallback

From Alignment to Actionable Stability

The theoretical ingredients behind SSCA

Multimodal alignment setting

$$H^{(i)} = f_i(X^{(i)}), \quad W_i \in \mathbb{R}^{p_i \times k}, \quad Z_0^{(i)} = T^{(i)}(H^{(i)})W_i$$

Goal: map multiple modalities into a shared k-dimensional subspace.

1. Deployment-shift model

$$\widetilde{H}^{(i)} = \phi^{(i)}(H^{(i)})$$

In-scope: coordinate-wise, orientation-preserving monotone distortions. Out-of-scope: non-monotone semantic shifts, heavy ties, and severe content changes.

2. Copula invariance & rank maps

$$F(x^{(1)}, \dots, x^{(d)}) = C(F_1(x^{(1)}), \dots, F_d(x^{(d)}))$$

Copula dependence is invariant under strictly increasing marginal transforms. Ranks and soft ranks approximate the marginals in finite samples.

3. Sliced-Wasserstein hub coupling

$$SW_2^2(P, Q) = \int_{\mathbb{S}^{k-1}} W_2^2(P_\theta, Q_\theta)\, d\theta$$

$$Q_B(u) = \sum_i w_i Q_i(u)$$

One-dimensional projections make transport efficient; the barycenter quantile

4. Spectral perturbation control

$$\|\sin \Theta(\hat{U}, U^*)\|_F \leq \frac{C_{\mathrm{DK}} \|E\|_2}{\gamma^*}$$

SSCA makes this bound actionable by decomposing E into rank, coupling, sampling, and numerical errors with measurable proxies.

These preliminaries motivate SSCA's three controls: clipped rank-probit stabilization, dependence-weighted hub coupling, and diagonal-stabilized spectral learning, plus a calibrated diagnostic gate.

Error decomposition

$$E = E_{\mathrm{rank}} + E_{\mathrm{cpl}} + E_{\mathrm{samp}} + E_{\mathrm{num}}$$

SSCA does not treat robustness as one black-box score; it makes each failure source measurable and actionable.

SSCA Deployment Protocol

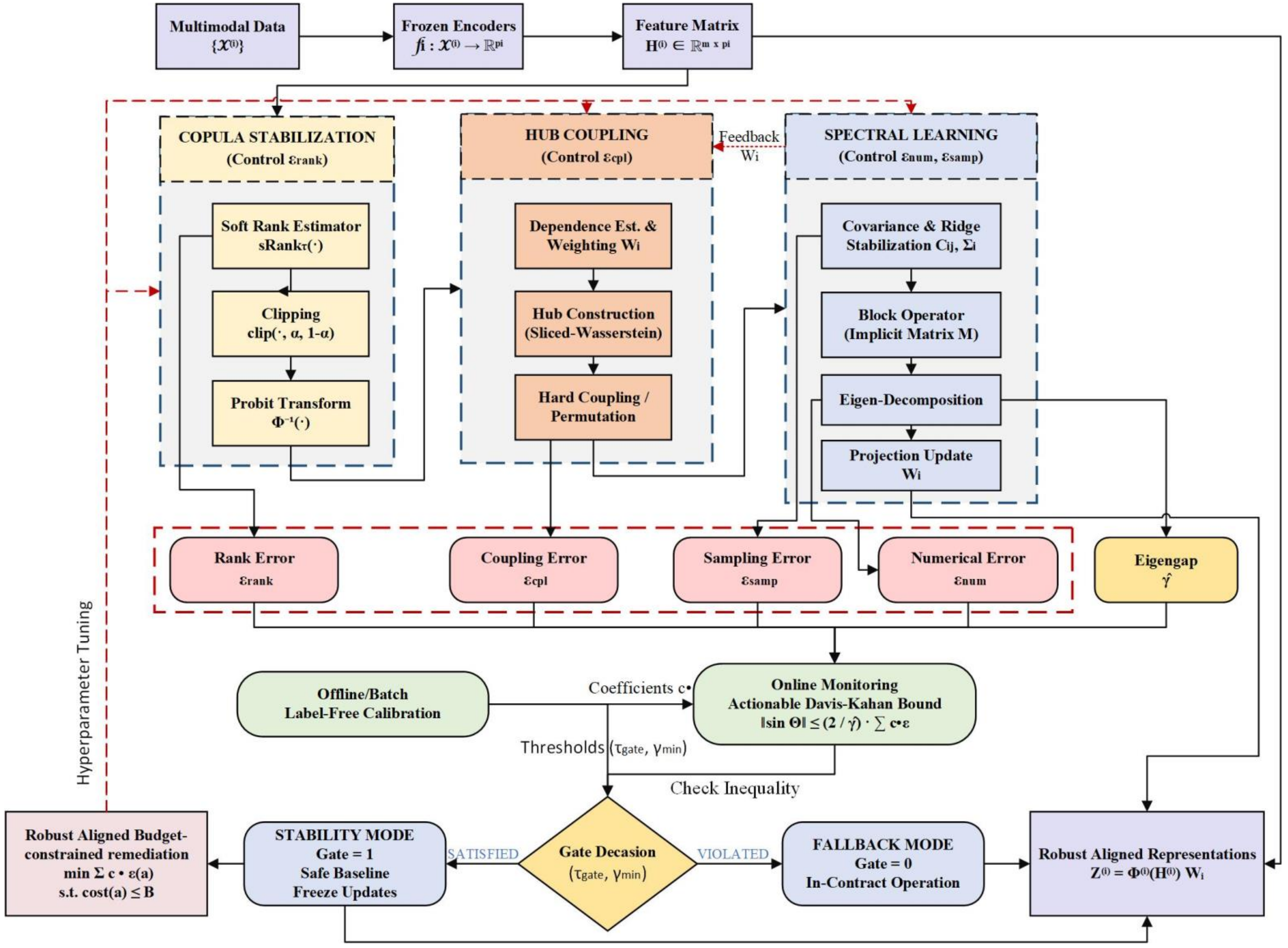

Figure 1. SSCA deployment protocol: three error-control modules control rank, coupling, sampling, and numerical errors, producing diagnostic proxies for the calibrated gate between Stability and Fallback Modes. The representation label in the monitoring path denotes the ideal shared subspace associated with U∗.

SSCA Protocol: Three Modules for Error Control

Approximating a copula-invariant dependence operator through controlled error sources.

1. Copula Stabilization—Control ε_{ran}

Goal: remove marginal sensitivity before alignment. For each modality and coordinate,

$$G_{:,j}^{(i)} = \Phi^{-1}(\mathrm{clip}(\mathrm{sRank}_{\tau_{rank}}(H_{:,j}^{(i)}), \alpha, 1-\alpha))$$

Clipping bounds the probit Lipschitz constant:

$$L_{\Phi-1}(\alpha) = \frac{1}{\varphi(\Phi^{-1}(\alpha))}, \qquad \alpha = 0.005 \Rightarrow L \approx 69.0$$

Finite-sample monotone invariance:

$$\left|\phi(g(x)) - \phi(x)\right| \le L_{\Phi-1}(\alpha)(2\varepsilon_t + \delta_{tie})$$

Rank proxy:

$$\hat{\varepsilon}_{rank} = \max_{i,j} \frac{1}{m} \left\| \mathrm{sRank}_{\tau_{rank}}(H_{:,j}^{(i)}) - \mathrm{uRank}(H_{:,j}^{(i)}) \right\|_1$$

2. Dependence-WeightedHubCoupling—Control ε_{cpl}

Goal: replace inconsistent pairwise maps by coherent multiway coupling. First compute, $Z_0^{(i)} = G^{(i)} W_i$

Projected Kendall dependence gives the modality weights:

$$\hat{\tau}_{ij} = \frac{1}{S_\tau} \sum_{s=1}^{S_\tau} \tau(Z_0^{(i)} \theta_s^\tau, Z_0^{(j)} \theta_s^\tau), \qquad \omega_{ij} \propto \exp([\hat{\tau}_{ij}]_+ / \tau_{dep}, \quad w_i \propto \sum_{j \neq i} [\hat{\tau}_{ij}\}_+.$$

For each sliced direction θ_s,

$$p_s^{(i)} = Z_0^{(i)} \theta_s, \quad \mathrm{Q}_s^{(B)}(u) = \sum_{i=1}^{d} w_i Q_s^{(i)}(u).$$

Hub discrepancy and coupling:

$$\beta_i = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{m} \sum_{t=1}^{m} (p_{s,(t)}^{(i)} - Q_s^{(B)}(u_t))^2, \quad u_t = \frac{t - \frac{1}{2}}{m},$$

$$i^* = \arg \min_i \beta_i, \quad \tilde{P}_i = \frac{1}{S} \sum \Pi_s^{(i \to i^*)}, \quad \Pi_i = \mathrm{Hungarian}(\tilde{P}_i), \quad \tilde{G}^{(i)} = \Pi_i G^{(i)}.$$

Coupling proxy:

$$\hat{\varepsilon}_{\text{rnd}} = \max_i \frac{\left\|\Pi_i - \tilde{P}_i\right\|_F}{\sqrt{m}}, \quad \hat{\varepsilon}_{\text{hub}} = \frac{\beta_{(2)} - \beta_{(1)}}{\beta_{(1)} + \varepsilon_{\text{hub}}}, \quad \hat{\varepsilon}_{\text{cpl}} = \hat{\varepsilon}_{\text{rnd}} + \frac{1}{\hat{\varepsilon}_{\text{hub}} + \varepsilon_{\text{hub}}} + c_{\text{tie}} \hat{\rho}_{\text{tie}}.$$

SSCA Protocol: Three Modules for Error Control

Approximating a copula-invariant dependence operator through controlled error sources.

3. Diagonal-Stabilized Spectral Learning — Control $\varepsilon_{\text{samp}}, \varepsilon_{\text{num}}$

Goal: learn an identifiable shared subspace with eigengap-aware stability. After coupling, center each modality:

$$\bar{G}^{(i)} = \tilde{G}^{(i)} - \frac{1}{m}\mathbf{1}\mathbf{1}^{\mathrm{T}}\tilde{G}^{(i)}.$$

Cross-covariances and stabilized whitening:

$$C_{ij} = \frac{1}{m}\bar{G}^{(i)\mathrm{T}}\bar{G}^{(j)}, \quad \Sigma_i = \frac{1}{m}\bar{G}^{(i)\mathrm{T}}\bar{G}^{(i)} + \lambda_{\text{stab}} I_{p_i}, \quad R_i = \Sigma_i^{-1/2}.$$

Block spectral operator:

$$M_{ij} = \omega_{ij} R_i C_{ij} R_j \ (i \neq j), \qquad M_{ii} = 0.$$

Equivalent spectral objective:

$$\max_{\{W_i^{\mathrm{T}}\Sigma_i W_i = I_k\}} \sum_{i<j} \omega_{ij}\,\mathrm{tr}(W_i^{\mathrm{T}} C_{ij} W_j) \Leftrightarrow \hat{U} = top-k \text{ eigenspace}(M), \quad W_i \leftarrow \mathrm{qf}(R_i \hat{U}_i).$$

Spectral proxies:

$$\hat{\varepsilon}_{\text{samp}} = \max_{i \neq j}\left\| C_{ij} - C_{ij}^{(\text{EMA})} \right\|_F, \quad \hat{\varepsilon}_{\text{num}} = \frac{\left\| M\hat{U} - \hat{U}\hat{\Lambda} \right\|_F}{\left\| \hat{U} \right\|_F}.$$

From Diagnostics to Deployment Actions

SSCA converts measurable error proxies into label-free update decisions.

Actionable Bound, Gate, and Fallback

SSCA turns diagnostics into a checkable subspace-risk bound:

$$\left\|\sin\Theta(\hat{U},U_s\right\|_F \le \frac{c_{\text{rank}}\hat{\varepsilon}_{\text{rank}} + c_{\text{clp}}\hat{\varepsilon}_{\text{clp}} + c_{\text{samp}}\hat{\varepsilon}_{\text{samp}} + c_{\text{num}}\hat{\varepsilon}_{\text{num}}}{\hat{\gamma}+10^{-8}}.$$

The coefficients $c_\bullet$ are deployment-local and calibrated from healthy unlabeled batches.

Stability gate: $$\text{Gate}(\tau_{\text{gate}},\gamma_{\min}) = \mathbf{I}\left[\frac{\sum_\bullet c_\bullet\hat{\varepsilon}_\bullet}{\hat{\gamma}+10^{-8}} \le \tau_{\text{gate}} \wedge \hat{\gamma} \ge \gamma_{\min}\right].$$

- Coefficients are calibrated from healthy unlabeled batches.
- The gate does not require target labels or the true population subspace.
- Every rejection is traceable to a diagnostic family and eigengap margin.
- When risk is high, SSCA suppresses unsupported alignment updates instead of forcing adaptation.

Gate = 1

Stability Mode

Allow alignment update under the calibrated stability envelope.

Gate = 0

Fallback Mode

Suppress unsafe update and use conservative no-update behavior.

Budgeted remediation

$$\min_{a\in A}\sum_\bullet c_\bullet\hat{\varepsilon}_\bullet(a) \quad \text{s.t.} \quad \text{cost}(a)\le B.$$

SSCA turns robustness into an auditable contract: update only when the dependence risk is bounded.

Experimental Results: Robustness under Shift

Frozen-feature evaluation across controlled distortions, retrieval stress tests, and deployment-style drifts.

Experimental setup

Datasets: CMU-MOSEI, MELD, MSCOCO, CC3M-500K
Protocol: frozen encoders; train projection heads and alignment parameters only
Perturbations: 10× affine scaling, JPEG compression, tokenizer and audio-codec changes
Baselines: contrastive, correlation-based, OT-based, and recent robust multimodal methods under the same frozen-feature / projection-budget protocol.

Metrics

$$\Delta(\pi) = s_{\text{clean}} - s_{\pi}$$

$$\text{Red}(\pi) = 100 \cdot \frac{\Delta_{\text{base}}(\pi) - \Delta_{\text{SSCA}}(\pi)}{\Delta_{\text{base}}(\pi)}$$

1. Controlled monotone distortion

10× affine scaling tests the formal monotone-drift contract. SSCA achieves the smallest degradation on both trimodal benchmarks.

1.8
CMU-MOSEI Δ↓ (vs. MMML 3.2)

1.5
MELD Δ↓ (vs. MMML 2.7)

≈44% lower degradation vs. MMML

2. Large-scale retrieval

CC3M-500K image–text retrieval with frozen CLIP ViT-B/16 under JPEG compression.

57.3
Recall@1 at JPEG Q=50

52.9%
degradation reduction at Q=50

Δ@Q=50: 2.4 for SSCA vs. 5.1 for CLIP baseline

3. Deployment-style drifts

JPEG compressor, tokenizer, and audio-codec changes are empirical stress tests beyond the exact invariance theorem.

1.6
average Δ↓ across six drift cases

2.2
MMML average Δ

SSCA reduces MELD Δ by 33% and CMU-MOSEI Δ by 28% vs. MMML.

SSCA consistently reduces degradation under controlled monotone shifts, frozen CLIP retrieval stress tests, and realistic raw-pipeline drifts.

Diagnostics, Ablation, and Operational Evidence

The diagnostic dictionary explains when SSCA should adapt, when it should fall back, and which module matters most.

1. Proxy bound verification

Synthetic ground-truth check compares the calibrated proxy bound with true subspace error.

0.967
Pearson r

0.941
Spearman ρ

95.0%
empirical coverage

2. Streaming drift detection and remediation

Gradual scaling drift on MELD: combined diagnostic detects upcoming degradation without labels.

0.94
AUROC

7
lead-time batches

11 / 13
batches to 95% recovery on MELD / CMU

Recovery is 45% / 43% faster than blind random adjustment.

3. Gate coverage and trade-offs

The calibrated gate keeps Stability Mode active for most episodes while limiting unsafe accepted updates.

12.5–18.9%
refuse rate

1.7–2.6%
unsafe accept rate

+0.8 to +1.2
net benefit

4. Ablation and efficiency profile

10× scaling degradation Δ↓ on CMU-MOSEI / MELD:

Full SSCA: 1.8 / 1.5 | w/o copula: 4.8 / 4.5
w/o hub: 3.8 / 3.5 | w/o spectral: 3.5 / 3.2

Removing copula stabilization causes the largest degradation increase, confirming that marginal stabilization is the dominant robustness factor.

Throughput at batch size 256: SSCA SoftRank **22.5k samples/s**; SSCA Sketch **24.8k samples/s**; MMML **19.2k samples/s**. Peak memory: **3.9–4.2 GB**.

SSCA enables **robust, auditable, safe-failure multimodal alignment** under deployment shift.
Robust: reduces marginal-sensitive alignment failure.
Auditable: makes rank, coupling, sampling, and numerical errors measurable.
Safe-failure: uses calibrated gating and fallback when adaptation is unsupported.

Thank You!