UniFast-HGR

Scalable and Efficient Maximal Correlation for Multimodal Models

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

Affiliation: Tsinghua University

Core

Make HGR maximal correlation usable for high-dimensional multimodal models.

A drop-in dependence regularizer that removes covariance whitening and preserves HGR-style alignment.

ICML 2026 | 43rd International Conference on Machine Learning | Seoul, South Korea, July 6-11, 2026

Motivation & Background : HGR is powerful, but not scalable

Why Modern Multimodal Models Need Scalable Dependence Learning

Motivation

Multimodal heterogeneity obscures true cross-modal dependencies: useful shared information is mixed with modality-specific artifacts, noise, scale mismatch, and long-tail effects.

scalable geometry

01 **Dot-product / distance / contrastive**

Efficient and compatible with large-scale pretraining such as InfoNCE and CLIP.

Mainly geometric; does not explicitly model nonlinear statistical dependence or directly target maximal-correlation dependence, which can matter under noise, long-tail distributions, or modality-specific artifacts.

statistical correlation

02 **CCA, Kernel CCA, Deep CCA**

Moves from linear correlation to nonlinear transformations and deep mappings.

But kernel/covariance storage, whitening, and eigensolves are costly and unstable.

neural HGR surrogates

03 **HGR, Soft-HGR**

HGR captures nonlinear dependence; Soft-HGR relaxes exact whitening.

But covariance-based variants still retain K×K interactions and $O(K^2)$ memory.

recent estimators

04 **CKA, dCor, IdCor**

Useful for representation similarity and nonlinear dependence analysis; IdCor adds intrinsic-dimension awareness.

But largely post-hoc or analysis-oriented; not ideal as scalable discriminative pretraining regularizers.

Practical gap Existing methods face a trade-off: scalable objectives often focus on geometric alignment, while statistically grounded dependence objectives remain computationally heavy.

Need: a stable, covariance-free, low-complexity maximal-correlation objective for ViTs, CLIP-style models, and high-dimensional multimodal foundation models.

Background Comparison: What Is Still Missing?

UniFast-HGR is motivated by the gap between scalable geometric alignment objectives and statistically grounded but computationally heavy dependence objectives.

Representative objectives and limitations

Family	Strength	Main limitation
Dot-product / contrastive	Efficient and scalable for large-scale representation learning	Mainly geometric; does not explicitly model nonlinear statistical dependence or directly target maximal-correlation dependence
CCA / KCCA / DCCA	Classical correlation view; nonlinear variants exist.	Whitening, covariance/kernel storage, and eigensolves limit scalability.
HGR / Soft-HGR / I-SoftHGR	Strong nonlinear dependence criterion.	Soft-HGR removes exact whitening but still uses K×K covariance; $O(K^2)$ memory.
CKA	Interpretable representation similarity.	Common variants rely on pairwise kernels; more suitable for representation analysis than online training
dCor / IdCor	Nonlinear dependence; IdCor considers intrinsic dimension.	dCor can be $O(N^2)$ and scale-sensitive; IdCor is mainly representation-comparison oriented.
MINE / NWJ / CLUB	Flexible neural MI estimation.	Variational or density-ratio critics introduce additional estimator design and optimization complexity

Motivation for UniFast-HGR

Preserve HGR-style dependence maximization

Remove covariance whitening and explicit K×K covariance storage.

Reformulation

centered cosine alignment + Gram-space structure + TSS diagonal suppression

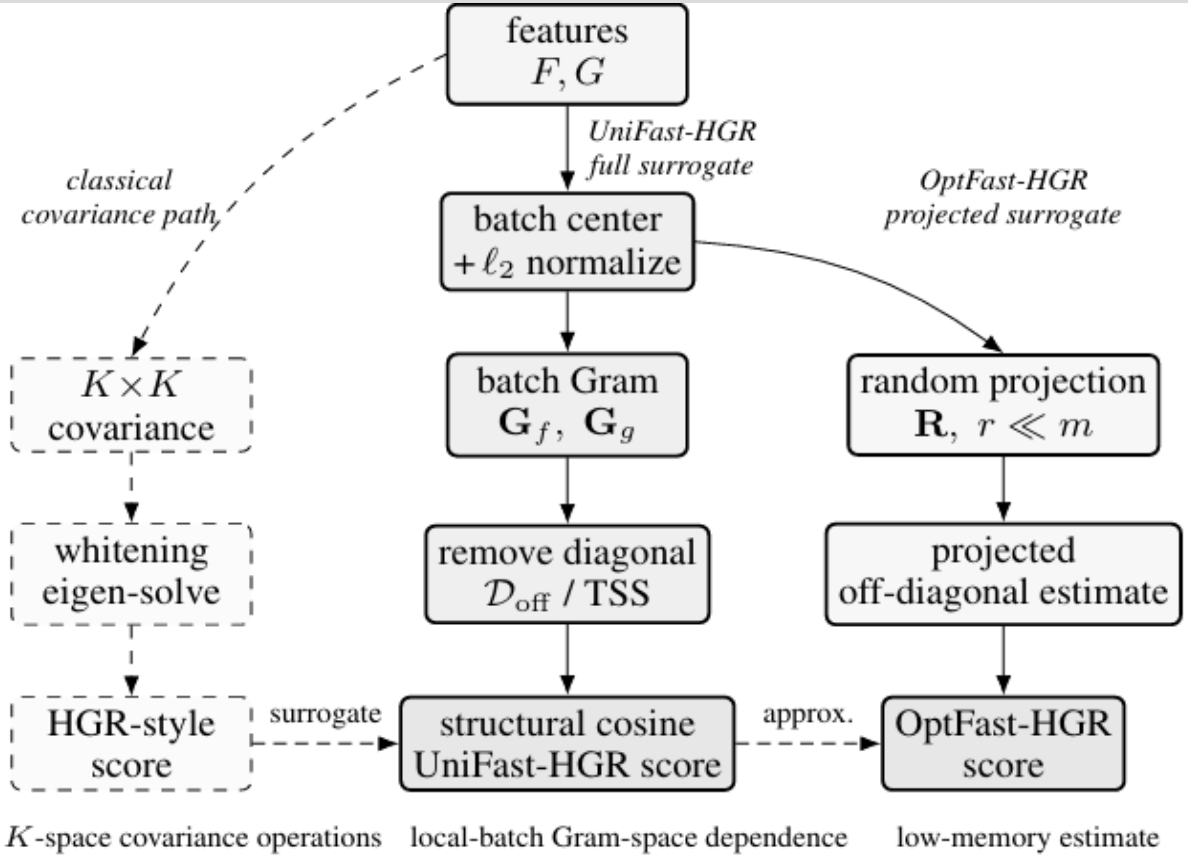
Scalability

UniFast-HGR: $O(m^2K)$
OptFast-HGR: $O(mKr)$

Takeaway

A plug-in dependence regularizer for contrastive, self-supervised, and supervised multimodal backbones.

Conceptual Computation Path



Path Summary.
Classical HGR follows a covariance-whitening path with $K \times K$ storage and eigensolve cost. UniFast-HGR moves the objective to centered, ℓ_2-normalized local Gram space, then uses TSS / (D_{off}) to focus on cross-sample structure. OptFast-HGR sketches the same off-diagonal signal for lower memory and faster large-batch training.

WHY IT WORKS — DESIGN PRINCIPLES

1. Variance control: Normalization bounds the loss, preventing magnitude-driven solutions.

2. Gram-space: Avoids $K \times K$ covariance storage via the trace bridge identity.

3. TSS: Removes invariant diagonal self-correlation, focuses on off-diagonal relations.

4. OptFast: Same structural signal at $O(mKr)$ via random projection.

5. Differentiable: Standard tensor ops; no architecture changes needed.

Proposed Method: From Covariance Whitening to Gram-Space Surrogate

Core idea: preserve maximal-correlation dependence learning while replacing whitening with bounded cosine and local-batch Gram

1) Classical path: covariance bottleneck

HGR maximizes dependence under strict whitening constraints.

Soft-HGR relaxes exact whitening, but still builds explicit feature covariance matrices.

$$\rho_{\mathrm{HGR}}(X,Y) = \sup_{\substack{f,g \\ \mathbb{E}[f]=\mathbb{E}[g]=0 \\ \mathrm{cov}(f)=\mathrm{cov}(g)=I}} \mathbb{E}\left[f(X)^\top g(Y)\right]$$

$$\mathcal{J}_{\mathrm{Soft}} = \mathbb{E}[f^\top g] - \tfrac{1}{2}\,\mathrm{tr}(\mathrm{cov}(f)\,\mathrm{cov}(g)), \quad \mathbb{E}[f] = \mathbb{E}[g] = 0$$

Limitation: whitening/eigensolve or explicit covariance storage becomes costly and unstable as K grows.

CCA/HGR: $O(K^3)$ compute
Soft-HGR: $O(K^2)$ memory

2) UniFast-HGR: normalized Gram-space surrogate

Replace finite-sample whitening with local-batch centering, standardization, row normalization, and Gram-space structural comparison.

Local-batch Standardization

$$\hat{f}_i = \frac{f_i - \mu_f}{\sigma_f + \epsilon}, \quad \hat{g}_i = \frac{g_i - \mu_g}{\sigma_g + \epsilon}$$

Row-wise ℓ_2 Normalization

$$\tilde{f}_i = \frac{\hat{f}_i}{\|\hat{f}_i\|_2 + \epsilon}, \quad \tilde{g}_i = \frac{\hat{g}_i}{\|\hat{g}_i\|_2 + \epsilon}$$

Paired Cosine Alignment

$$\langle \tilde{f}_i, \tilde{g}_i \rangle = \frac{\hat{f}_i^\top \hat{g}_i}{\|\hat{f}_i\|_2 \|\hat{g}_i\|_2} = \cos(\hat{f}_i, \hat{g}_i)$$

Gram-space Structural Dependence

$$G_f = \tilde{F}\tilde{F}^\top, \quad G_g = \tilde{G}\tilde{G}^\top$$

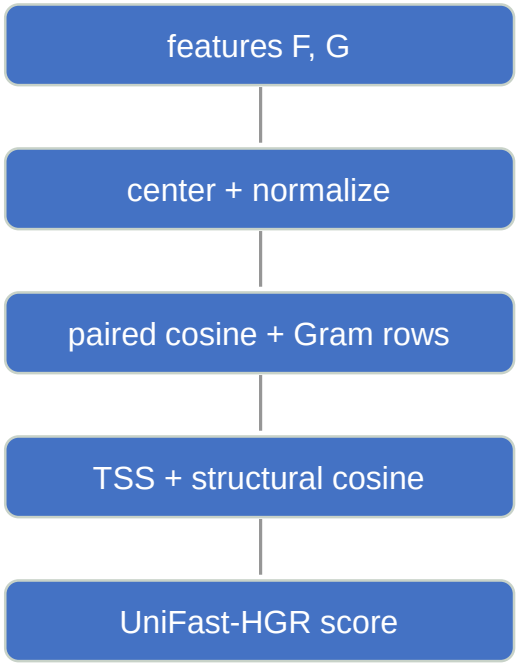
$$\mathrm{tr}\left(\tilde{F}^\top \tilde{F} \tilde{G}^\top \tilde{G}\right) = \mathrm{tr}\left(\tilde{F}\tilde{F}^\top \tilde{G}\tilde{G}^\top\right) = \langle G_f, G_g \rangle_F$$

$$\mathrm{trace} \approx \frac{1}{2(m-1)} \sum_{i=1}^{m} \cos\left((G_f)_{i:}, (G_g)_{i:}\right)$$

Effect: estimate covariance interaction through sample relations instead of explicit K × K covariance matrices.

3) Plug-in auxiliary objective

The encoder, task loss, and downstream protocol remain unchanged. UniFast-HGR is added as a differentiable regularizer on latent features.



UniFast-HGR Objective: Off-Diagonal Structural Dependence

The objective combines paired cosine alignment with structural alignment over off-diagonal Gram rows.

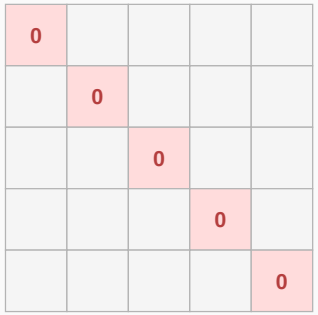
Why TSS is needed

After row-wise normalization, the Gram diagonal is nearly fixed: it is self-correlation, not cross-sample dependence. If retained, it can dominate the structural term in small or homogeneous batches.

$$(G_f)_{ii} = \langle \bar{f}_i, \bar{f}_i \rangle = \|\bar{f}_i\|_2^2 \approx 1 \qquad (G_g)_{ii} = \langle \bar{g}_i, \bar{g}_i \rangle = \|\bar{g}_i\|_2^2 \approx 1$$

Trivial Spectrum Suppression

$$D_{\text{off}}(G) = G - \text{Diag}(\text{diag}(G)).$$



diagonal removed; only cross-sample relations remain

Final UniFast-HGR objective

Instance-Level Alignment

$$J_{align} = \frac{1}{N-1}\sum_{i=1}^{N} \cos(\bar{f}_i, \bar{g}_i)$$

Off-Diagonal Structural Term

$$J_{struct} = \frac{1}{N-1}\sum_{i=1}^{N} \cos((D_{off}(G_f))_{i:}, (D_{off}(G_g))_{i:})$$

UniFast-HGR Objective

$$J_{UniFast} = J_{align} - \frac{\lambda}{2} J_{struct}$$

$$J_{UniFast} = \frac{1}{N-1}\sum_{i=1}^{N} \cos(\bar{f}_i, \bar{g}_i) - \frac{\lambda}{2(N-1)}\sum_{i=1}^{N} \cos((D_{off}(G_f))_{i:}, (D_{off}(G_g))_{i:})$$

Both terms are bounded by normalization, improving numerical stability.

The structural term compares local-batch relational distributions rather than feature covariance matrices.

The objective is a standard differentiable auxiliary loss added to the task loss.

Multimodal Generalization, OptFast-HGR, and Complexity

UniFast-HGR scales from two modalities to multiple modalities and uses OptFast-HGR for low-memory structural estimation.

Multimodal UniFast-HGR: pairwise aggregation

For M modalities, aggregate pairwise maximal-correlation surrogates. This is a practical pairwise framework, not a closed-form higher-order HGR theory.

$$J_{Mnlti} = \sum_{1\le u<v\le M} \left[\frac{1}{N-1}\sum_i \cos(f_i^{(u)}, f_i^{(v)}) - \frac{\lambda}{2(N-1)}\sum_i \cos((D_{off}(\text{distri}_i^{(u)}), D_{off}(\text{distri}_i^{(v)}))\right]$$

Uniform pair weights are used by default; task-guided weights can be added without changing the objective.

Cost grows linearly with the number of modality pairs in Gram space.

OptFast-HGR: projected off-diagonal estimate

For large local batches, avoid materializing the full m × m Gram structure by projecting the off-diagonal signal into a low-dimensional sketch.

$$F \in \mathbb{R}^{m\times K}, \qquad R \in \mathbb{R}^{m\times r}, \qquad r \ll m.$$

$$\widehat{G}_f^{\text{proj}} = D_{\text{off}}\left(FF^\top\right) R.$$

$$\widehat{G}_f^{\text{proj}} = F(F^\top R) - \text{Diag}(d_F)R, \qquad d_F = \text{diag}\left(FF^\top\right).$$

$$\text{For row-normalized features:} \qquad d_F \approx \mathbf{1}.$$

The same projection is applied to the paired modality before structural cosine computation.

Complexity and scalability summary

Method family	Main operation	Cost / memory	Scalability implication
Classical CCA/HGR	whitening / eigensolve	$O(K^3)$	intractable as K grows
Soft-HGR	K x K covariance interaction	$O(K^2)$ memory	OOM risk at high feature dimensions
UniFast-HGR	local-batch Gram structure	$O(m^2K)$; Gram memory $O(m^2)$	feature storage remains linear in K
OptFast-HGR	random projected off-diagonal sketch	$O(mKr)$; sketch memory $O(mr)$	low-memory approximation for large local batches

Microbenchmark: at $K = 10^5$, Soft-HGR is OOM on a 24GB GPU, while UniFast-HGR runs stably with 1.55GB VRAM.

Complexity and End-to-End Scalability

UniFast-HGR removes the covariance whitening bottleneck while keeping feature storage linear in the feature dimension K.

Complexity principle

Local-batch Gram operator

m = local per-device batch size
K = feature dimension
r = projection dimension

CCA / HGR
$O(K^3)$ compute

Soft-HGR
$O(K^2)$ memory

UniFast-HGR
$O(m^2K)$, Gram $O(m^2)$

OptFast-HGR
$O(mKr)$, sketch $O(mr)$

Key distinction

The global batch can be large, while the local feature block used for the Gram operator remains moderate. This makes the memory accounting compatible with data-parallel training.

Scalability microbenchmark at extreme feature dimensions

Synthetic features, local batch size m = 256, single 24GB GPU. OOM: out of memory (>24GB).

Method	K	Time ↓	VRAM ↓	Status
Soft-HGR	10^3	1.80s	0.98GB	Stable
Soft-HGR	10^4	95.6s	15.3GB	High cost
Soft-HGR	10^5	—	>24GB	**OOM**
CCA	10^3	12.4s	2.50GB	Slow
CCA	10^4	—	>24GB	**OOM**
UniFast-HGR	10^3	0.07s	0.82GB	**Stable**
UniFast-HGR	10^4	0.71s	1.12GB	**Stable**
UniFast-HGR	10^5	**7.20s**	**1.55GB**	**Stable**
InfoNCE ref.	10^5	6.85s	1.45GB	Reference
VICReg ref.	10^5	7.15s	1.62GB	Reference
Barlow Twins ref.	10^5	7.28s	1.68GB	Reference

Takeaway: UniFast-HGR is stable at $K = 10^5$ with 1.55GB VRAM, while Soft-HGR is OOM; its efficiency profile is close to lightweight SSL references.

Experimental Evidence: Accuracy and Robustness

Evaluation covers foundation-model tuning, retrieval, remote sensing, emotion recognition, and ablation.

Large-scale foundation-model tuning

ImageNet-1K Top-1 Accuracy

Backbone	Baseline	UniFast	Gain
CLIP	76.1	**80.4**	+4.3
SigLIP	81.3	**84.8**	+3.5
DINOv2	81.8	**85.3**	+3.5

COCO Text-Image R@1

Backbone	Base	UniFast
CLIP	38.9	**42.1**
SigLIP	50.8	**53.8**
DINOv2	51.1	**53.9**

Video-text retrieval (ViCLIP T2V R@1): MSR-VTT 36.4 → 43.3, LSMDC 17.1 → 20.7, DiDeMo 16.4 → 20.5.

Downstream multimodal tasks

Remote sensing classification: Berlin

Method	OA	Time/epoch
Dot product	75.20	23.18s
Soft-HGR	65.80	25.83s
UniFast-HGR	**80.75**	24.53s
OptFast-HGR	**80.46**	23.54s

Remote sensing segmentation

Dataset	OA	mIoU
Vaihingen	**93.01**	**84.62**
Globe230k	**91.48**	**76.36**

IEMOCAP ACC: no missing 71.29 → 73.66; Text+Audio 67.85 → 70.94; 80% labels hidden 57.75 → 62.05.

Mechanism analysis and overall claims

Component-wise ablation

Variant	Berlin OA	ImageNet Top-1
Soft-HGR	65.80	76.3
+ cosine	75.23	77.8
+ variance	78.65	79.2
+ structure	80.62	79.8
+ TSS / UniFast	**80.75**	**80.1**
+ OptFast	80.46	79.6

Scalability

Integrates into ViT / CLIP-style backbones without changing encoders or task protocols.

Accuracy

Improves over baselines, Soft-HGR, CKA, dCor, and IdCor across classification and retrieval.

Robustness

TSS reduces degradation under small, skewed, or homogeneous batches and missing/limited labels.

Takeaway

UniFast-HGR provides a lighter memory and runtime profile while strengthening multimodal backbones.

Thank You!