InfoNCE and Beyond: The Evolution of Multimodal Representation Learning

Prateek Khanna

SolFinder Research

Abstract

Contrastive learning has become the dominant paradigm for aligning heterogeneous modalities in multimodal large language models (MLLMs). Information Noise Contrastive Estimation (InfoNCE) Loss and its variants serve as the foundational training objectives for such models. As models scale to billions of parameters and training batches reach millions of samples, the limitations of conventional softmax-based contrastive losses have become increasingly apparent: quadratic memory growth, batch size sensitivity, semantic noise in negative sampling, and the inherent tension between alignment and uniformity. This paper presents a comprehensive survey of the evolution from InfoNCE to next-generation objectives for multimodal LLMs. We analyze the theoretical foundations of contrastive learning through the alignment-uniformity framework, examine the practical innovations that have enabled unprecedented batch scaling, and evaluate emerging alternatives including sigmoid-based losses, non-contrastive alignment methods, and hybrid generative-discriminative objectives. The field is undergoing a fundamental shift from batch-dependent contrastive paradigms toward more scalable, stable, and theoretically grounded objectives. We conclude with a research agenda addressing the open challenges in loss design for the next generation of multimodal foundation models.

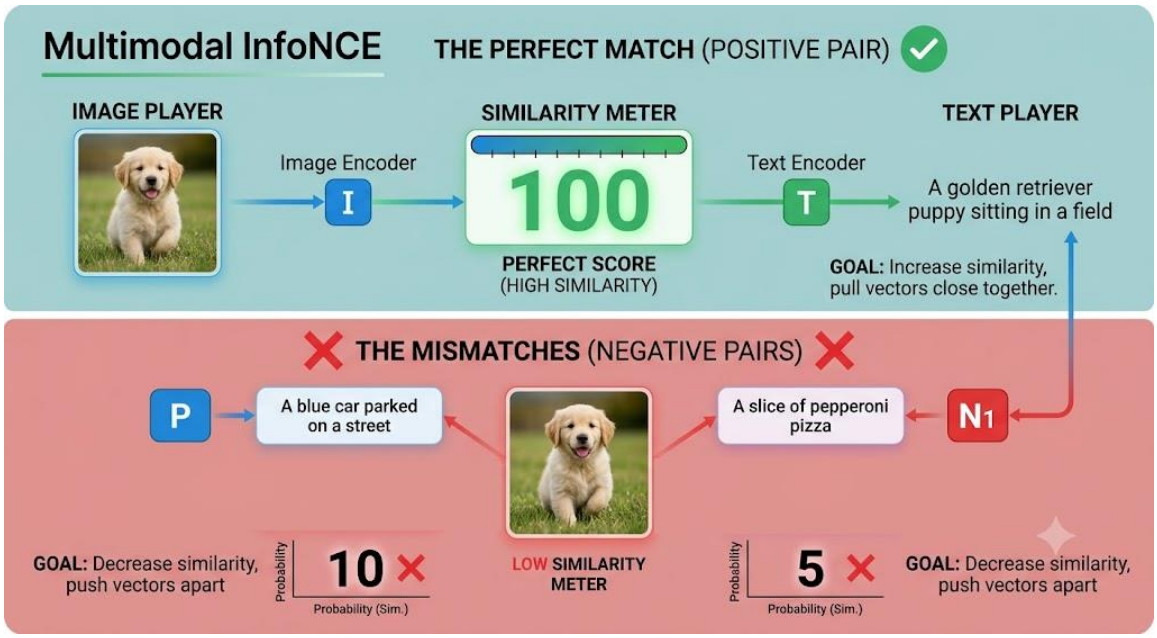
Keywords: Contrastive Learning, InfoNCE, Multimodal LLMs, SigLIP, Non-Contrastive Learning, Alignment-Uniformity, Vision-Language Models, Representation Learning

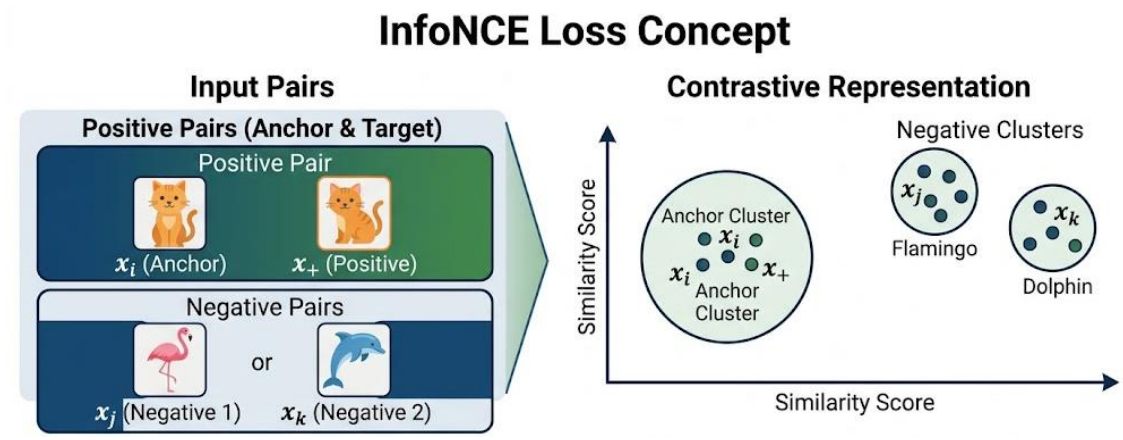
Introduction

The Contrastive Revolution

The emergence of CLIP (Contrastive Language-Image Pre-training) in 2021 marked a watershed moment in artificial intelligence. By training dual encoders - vision and language - on 400 million image-text pairs using a simple contrastive objective, CLIP demonstrated that natural language could serve as a scalable supervisory signal for visual understanding. The model achieved remarkable zero-shot classification, cross-modal retrieval, and transfer learning capabilities, establishing a template that has been replicated across virtually all subsequent multimodal systems.

At the heart of CLIP's success lies the InfoNCE loss - a contrastive objective that pushes representations of paired modalities closer together while repelling unpaired combinations. This approach, rooted in noise-contrastive estimation, has proven extraordinarily effective at learning shared embedding spaces where semantic similarity is preserved across modalities. The success of CLIP spawned an entire ecosystem of multimodal models: LLaVA, Qwen-VL, Stable Diffusion, DALL-E, and countless others all build upon or extend its contrastive foundation.





However, as the field has matured, the limitations of InfoNCE have also become increasingly apparent. The loss requires large batch sizes to provide sufficient negative samples, exhibits quadratic memory scaling, is sensitive to hyperparameter choices (particularly temperature), and struggles with semantic noise in web-crawled training

data. Most critically, the assumption that all non-paired samples are valid negatives breaks down in real-world datasets where images may be partially relevant to multiple captions within a batch.

The Scaling Imperative

The contemporary landscape of multimodal AI is defined by scale. State-of-the-art models now process billions of parameters, train on trillion-token datasets, and operate across increasingly diverse modality combinations - vision, language, audio, video, structured data, and embodied signals. This scaling has placed unprecedented demands on training objectives:

Batch size scaling: Cheng. Z et al. (2024) demonstrated tile-based computation strategies that enable CLIP training with batch sizes of 4 million or 12 million samples, achieving 78× memory reduction at 256K batch size and 281× at 1M batch size. Yet even these innovations reveal that "excessively large batch sizes...result in suboptimal performance," suggesting fundamental limits to the contrastive paradigm.

Semantic fidelity: Web-crawled datasets contain loose alignments where a single image may be partially relevant to multiple captions. The core assumption of InfoNCE - that each positive pair is surrounded by mutually exclusive negatives—is systematically violated, introducing noisy supervision that impairs representation quality.

Multi-objective training: Modern multimodal LLMs increasingly blend contrastive pre-training with captioning, masked prediction, self-distillation, and generative objectives. The standalone contrastive paradigm is giving way to hybrid approaches that demand more flexible loss formulations.

Research Objectives

This paper pursues three objectives: (1) to provide a rigorous theoretical analysis of InfoNCE and its fundamental properties; (2) to survey the landscape of innovations that have extended, modified, or replaced InfoNCE in contemporary multimodal systems; and (3) to identify the emerging paradigms that may define the next generation of multimodal representation learning.

Theoretical Foundations: InfoNCE and the Alignment-Uniformity Framework

The InfoNCE Objective

The canonical InfoNCE loss for multimodal contrastive learning takes the symmetric form:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2K}\sum_{i=1}^{K}\left[\log\frac{e^{\langle f_x(x_i), f_t(t_i)\rangle/\tau}}{\sum_{j=1}^{K} e^{\langle f_x(x_i), f_t(t_j)\rangle/\tau}} + \log\frac{e^{\langle f_x(x_i), f_t(t_i)\rangle/\tau}}{\sum_{j=1}^{K} e^{\langle f_x(x_j), f_t(t_i)\rangle/\tau}}\right]$$

where f_x and f_t are modality-specific encoders, τ is a temperature parameter, and K is the batch size. The loss operates by treating each correct image-text pair as a positive and all other combinations within the batch as negatives, using a cross-entropy formulation over the similarity matrix.

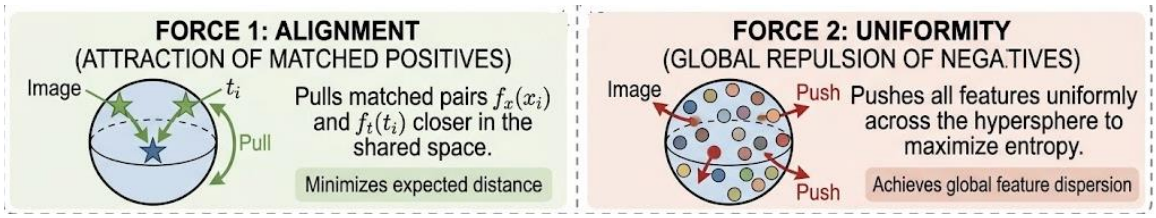
Alignment and Uniformity

Wang, T., & Isola, P. (2020, November) established the foundational theoretical understanding of contrastive learning by identifying two key properties that the loss optimizes:

Alignment: Features from positive pairs should be close together in the embedding space. This is measured by the expected distance between paired samples: $\mathcal{L}_{\text{align}}(f;\alpha) = -\mathbb{E}_{(x,y)\sim p_{\text{pos}}}[\|f(x) - f(y)\|_2^{\alpha}]$
Uniformity: Features should be uniformly distributed on the unit hypersphere, preserving maximal information. This is quantified by the Gaussian potential: $\mathcal{L}_{\text{uniform}}(f;t) = \log \mathbb{E}_{x,y\sim p_{\text{data}}}[e^{-t\|f(x)-f(y)\|_2^2}]$
Critically, Wang and Isola proved that as the number of negative samples M → ∞, the normalized contrastive loss converges to:

$$\lim_{M\to\infty} \mathcal{L}_{\text{contrastive}}(f;\tau,M) - \log M = -\frac{1}{\tau}\mathbb{E}_{(x,y)\sim p_{\text{pos}}}[f(x)^T f(y)] + \mathbb{E}_{x\sim p_{\text{data}}}[\log \mathbb{E}_{x^-\sim p_{\text{data}}}[e^{f(x)^T f(x^-)/\tau}]]$$

The first term encourages alignment; the second term, through its log-sum-exp form, pushes all features apart, promoting uniformity. This decomposition reveals that contrastive learning is fundamentally optimizing a trade-off: bringing positives together while maintaining global feature dispersion.



Theoretical Limitations

The alignment-uniformity framework exposes several inherent tensions:

The finite-negative gap: In practice, M is finite (bounded by batch size), meaning the loss only approximates the ideal alignment-uniformity objective. Directly optimizing L_align and L_uniform can outperform InfoNCE with finite negatives.
Temperature sensitivity: The temperature τ controls the sharpness of the distribution.

Low temperatures emphasize hard negatives but can cause collapse; high temperatures soften the objective but reduce discriminative power. Finding optimal τ requires extensive tuning.
The mutual exclusivity assumption: InfoNCE assumes that all non-paired samples are valid negatives. In multimodal datasets, this assumption is systematically violated: an image of a "dog playing in a park" may be semantically related to multiple captions in a batch ("canine outdoors," "pet exercise," "grass field animal").

Core Limitation	**Technical Phenomenon**	**Impact on Scaling / Performance**
Quadratic Memory Growth	Full materialization of the $K \times K$ similarity matrix and its batch-wide softmax normalization.	Constrains practical training batch sizes to approximately 32K–64K unless sophisticated tiling strategies are used.
Batch Size Sensitivity	Performance is highly dependent on large batches to supply enough negative samples; however, excessively large batches can introduce harmful hard negatives.	Creates extensive hyperparameter tuning demands (especially around the temperature parameter) and can cause geometric embedding collapse or suboptimal representation quality.
Mutual Exclusivity Assumption	The mathematical structure assumes that every non-paired sample in a batch is a strictly invalid negative.	Systematically violated by web-crawled datasets containing semantic looseness (e.g., cross-relevant text) or semantic redundancy, introducing noisy supervision.
Multi-Objective Incompatibility	Softmax normalization natively interacts and clashes with other loss terms during joint optimization.	Makes it technically difficult to cleanly blend standalone contrastive learning with generative, captioning, or self-supervised objectives.

Innovations in Contrastive Learning for Multimodal LLMs

Scaling Batch Sizes: The Memory Challenge

The effectiveness of contrastive learning is strongly correlated with batch size - larger batches provide more negative samples, improving representation quality. However, the full materialization of the K × K similarity matrix in InfoNCE creates quadratic memory growth, historically limiting practical batch sizes.
Inf-CLIP (Cheng Z. et al., 2024) introduced a tile-based computation strategy that partitions the contrastive loss calculation into arbitrary small blocks, avoiding full materialization of the similarity matrix. Combined with a multi-level tiling strategy leveraging hierarchical distributed systems (ring-based GPU communication and fused CUDA kernels) this approach achieved unprecedented scalability:

- Batch sizes of 4M with 8 A800 GPUs or 12M with 32 GPUs
- 78× memory reduction at 256K batch size
- 281× memory reduction at 1M batch size (128 GPUs)
- No accuracy degradation compared to full-matrix computation

However, the authors observed that "excessively large batch sizes result in suboptimal performance," suggesting that beyond a certain point, additional negatives provide diminishing returns or introduce harmful hard negatives that distort the embedding geometry.

Hard Negative Mining

The quality of negatives matters as much as their quantity. Naive random sampling within a batch often yields easy negatives that provide little training signal. Hard negative mining strategies aim to identify and emphasize the most informative negative samples:

Adaptive and Adversarial Sampling: Methods like Adversarial Contrastive Estimation (ACE) employ generator networks that actively search for negatives maximizing the loss, creating a minimax game between the encoder and negative generator (Bose A.J. et al., 2018,July). Regularization (e.g., entropy constraints) prevents mode collapse of the negative distribution.
Hard Negative Synthesis: When natural hard negatives are scarce, synthetic generation creates more challenging training examples. MoCHi and SSCL generate hard negatives through linear transformation or convex mixing of existing samples (Kalantidis, Y. et al., 2020). DropMix extends this to partial-dimension mixing, focusing on a fraction of feature dimensions to synthesize harder negatives without information collapse.

Scalable Retrieval: For large-scale datasets, approximate nearest-neighbor (ANN) algorithms enable global hard negative mining. Locality-Sensitive Hashing (LSH) converts continuous embeddings into binarized codes, enabling sublinear retrieval of hard negatives across millions of samples with minimal recall loss.
Curriculum Learning: Training stability is improved by adapting negative hardness over epochs. Ring-based conditional sampling narrows the similarity band as embedding geometry matures, following a curriculum that avoids the gradient explosion associated with always selecting the single hardest negative.

Strategy Type	**Specific Methods Mentioned**	**Core Mechanism / Approach**	**Target Outcome**
Adaptive & Adversarial	Adversarial Contrastive Estimation (ACE)	Employs generator networks that actively search for loss-maximizing negatives in a minimax game with the encoder, bound by entropy constraints.	Maximizes informative gradient signals while preventing mode collapse of the negative distribution.
Synthetic Generation	MoCHi, SSCL, DropMix	Synthesizes hard negatives through linear transformation, convex mixing, or partial-dimension mixing of existing samples in the batch.	Generates challenging negative samples on-the-fly when natural hard negatives are scarce, avoiding information collapse.
Scalable Retrieval	Locality-Sensitive Hashing (LSH) via ANN algorithms	Converts continuous embeddings into binarized codes to perform sublinear global hard negative lookups across millions of samples.	Makes global hard negative mining computationally viable for massive datasets with minimal recall loss.

Strategy Type	Specific Methods Mentioned	Core Mechanism / Approach	Target Outcome
Curriculum Learning	Ring-based conditional sampling	Narrows the similarity band of selected negative samples over epochs as embedding geometry matures.	Improves training stability by avoiding early-stage gradient explosions caused by excessively hard negatives.

Addressing Semantic Noise: The CLIPin Approach

CLIPin (Yang, S. et al., 2025) directly addresses the semantic noise problem in InfoNCE. The core insight is that web-crawled datasets contain two types of semantic violations:

Semantic looseness: A single image or caption may be partially relevant to multiple samples within a batch (common in natural image-text datasets)
Semantic redundancy: Semantically similar samples are treated as negative pairs (common in medical datasets with limited description variety)

CLIPin introduces a non-contrastive plug-in that enhances instance-level semantic alignment without abandoning the contrastive framework. By modeling fine-grained semantic correspondence, it mitigates the noisy supervision that degrades representation quality under InfoNCE's mutual exclusivity assumption.

Beyond InfoNCE: Next-Generation Objectives

Sigmoid-Based Losses: The SigLIP Revolution

The most significant departure from InfoNCE in recent years is the adoption of sigmoid-based contrastive losses. SigLIP (Sigmoid Loss for Language-Image Pre-training), developed by Google DeepMind, replaces the softmax normalization over the batch with independent per-pair sigmoid functions:

$$\mathcal{L}_{\text{SigLIP}} = -\sum_{i=1}^{K}\sum_{j=1}^{K}\left[\mathbb{1}_{i=j}\log\sigma\left(\frac{\langle u_i, v_j\rangle}{\tau}+b\right)+\mathbb{1}_{i\neq j}\log\left(1-\sigma\left(\frac{\langle u_i, v_j\rangle}{\tau}+b\right)\right)\right]$$

where σ is the sigmoid function and b is a learned bias parameter.
(Zhai, X. et al., 2023).
Key advantages over InfoNCE:

Memory efficiency: The sigmoid loss treats each image-text pair independently, eliminating the need to compute the full $K \times K$ similarity matrix and its softmax normalization. This enables scaling to much larger batch sizes without the memory overhead of the full softmax matrix.
Scalable training: SigLIP's independence from batch-wide normalization means that batch size can be increased without the quadratic memory penalty that constrains InfoNCE-based methods.
Theoretical characterization: The global minima of the SigLIP loss are characterized by (m,b)-Constellations—collections of embeddings satisfying:
- $\langle U_i, V_i \rangle \geq m + b$ (positive pairs separated by margin)
- $\langle U_i, V_j \rangle \leq -m + b$ for all $i \neq j$ (negative pairs pushed below threshold)

This provides explicit geometric control over the embedding space, with dimension-cardinality bounds determined by spherical code analogies.
SigLIP 2 (2025) extends this foundation by blending the sigmoid contrastive objective with decoder-based captioning and localization losses (cross-entropy for text, L2 for bounding boxes), local-to-global self-distillation and masked-prediction consistency, multi-objective training with loss scheduling, multilingual support with de-biasing techniques, and NaFlex variant preserving native aspect ratios with multi-resolution training. SigLIP models now serve as the vision encoder backbone in Google's PaliGemma VLM family, demonstrating production-scale validation of the sigmoid approach.

Non-Contrastive Alignment: The NOVA Framework

A more radical departure from contrastive learning is represented by NOVA (NOn-contrastive Vision-language Alignment), which eliminates negative samples entirely. NOVA aligns visual representations to a frozen, domain-specific text encoder by predicting text embeddings from augmented image views, while enforcing an isotropic Gaussian structure via Sketched Isotropic Gaussian Regularization (SIGReg).(Kuhn, L. et al., 2026)

Key innovations of this approach include:

No negative sampling: Eliminates the need for momentum encoders, stop-gradients, or large batch sizes
Single hyperparameter: The training objective reduces to a single tunable parameter (SIGReg strength)
Distributional regularization: Enforces isotropic Gaussian structure on embeddings, preventing dimensional collapse

Superior stability: Exhibits substantially more consistent training runs across random seeds

On zero-shot chest X-ray classification using ClinicalBERT as the text encoder, NOVA outperforms multiple standard contrastive baselines while requiring significantly less hyperparameter tuning. This demonstrates that non-contrastive vision-language pretraining offers a simpler, more stable, and more effective alternative to contrastive methods in certain domains.

Relaxing Contrastiveness: The ReCo Approach

ReCo (Relaxing Contrastiveness) takes an intermediate position, modifying the InfoNCE energy function without abandoning the contrastive framework. By adjusting the contrastive strength parameter, ReCo achieves consistent improvements in both image-image and text-image retrieval (Lin, Z. et al., 2023). On the CheXpert 8×200 retrieval dataset, ReCo is found to improve InfoNCE by 2.9% in average retrieval precision using identical architecture and training protocols. This demonstrates that the loss formulation itself, independent of architectural changes, provides meaningful scope for improvement.

LLM-Enhanced Contrastive Learning: LLM2CLIP

The emergence of powerful large language models has created new opportunities for enhancing multimodal representation learning. LLM2CLIP leverages LLMs' advanced text understanding and open-world knowledge to improve CLIP's ability to process long and complex captions (Huang, W. et al., 2024).
Technical approach:
Caption contrastive fine-tuning: The LLM is fine-tuned with contrastive learning in the caption space, improving the discriminability of its output embeddings
Efficient training: The LLM is frozen during CLIP training; only a lightweight adapter, the vision encoder, and two projectors are updated
Pre-extracted features: Text features are pre-extracted and stored, ensuring training memory and computational costs remain nearly identical to standard CLIP

Despite minimal overhead, LLM2CLIP achieves transformative improvements in downstream tasks including long and short text retrieval, cross-language retrieval, and LLaVA training. The key insight is that LLMs' textual capabilities can be "extracted" into the embedding space, providing richer, higher-dimensional supervision than vanilla text encoders.

Multi-Objective and Hybrid Approaches

Contemporary state-of-the-art multimodal models increasingly blend multiple objectives rather than relying on contrastive learning alone:
SigLIP 2's hybrid training combines sigmoid contrastive loss for global alignment, decoder-based captioning loss for fine-grained text generation, self-distillation for consistency across resolutions, and masked prediction for local feature learning.
MMCL (Multimodal Contrastive Learning) framework integrates intra-modal and inter-modal contrastive losses, geometric multi-view formulations (GMC), synergy and unique information discovery via Partial Information Decomposition (CoMM), and Mixup-contrastive losses for many-to-many alignment (M3CoL).
Generative-augmented contrastive learning uses GAIMG-CL to preprocess and refine noisy textual data before contrastive training, while parallel contrastive learning decomposes the learning process into specialized stages for item and user representations.

Objective / Framework	**Core Technical Mechanism**	**Primary Innovations & Strengths**	**Domain / Model Use Case**
SigLIP / SigLIP 2 *(Sigmoid-Based)*	Replaces batch softmax with independent, pairwise sigmoid functions using learned bias b and margin m (m,b) - Constellations).	Linear memory scaling, enables batch sizes >1M without batch-wide stats, and cleanly integrates with multi-objective scheduling (captioning, masked prediction).	General vision-language pre-training; acts as the visual backbone for Google's PaliGemma VLM family.
NOVA *(Non-Contrastive Framework)*	Predicts text embeddings from augmented image views using a frozen text encoder, regularized by Sketched	Completely eliminates negative samples, momentum encoders, and stop-gradients; reduces tuning to a single	Specialized domains with highly redundant/noisy negative structures (e.g., zero-shot chest

Objective / Framework	Core Technical Mechanism	Primary Innovations & Strengths	Domain / Model Use Case
	Isotropic Gaussian Regularization (SIGReg).	hyperparameter while preventing dimensional collapse.	X-ray classification).
ReCo *(Relaxed Contrastiveness)*	Modifies the InfoNCE energy function by adjusting the contrastive strength parameter without abandoning the framework.	Enhances retrieval precision solely through loss formulation adjustments, independent of any architecture changes.	Image-to-image and text-to-image retrieval tasks (e.g., validated on CheXpert retrieval).
LLM2CLIP *(LLM-Enhanced Alignment)*	Captions are fine-tuned with contrastive learning in text space using a frozen LLM, connecting to the vision encoder via an adapter and projectors.	Distills an LLM's advanced open-world text comprehension into the multimodal embedding space with virtually zero computational overhead during CLIP training.	Enhancing long/complex caption processing, cross-lingual retrieval, and boosting LLaVA training performance.

Comparative Analysis and Emerging Patterns

The Sigmoid-Softmax Dichotomy

The transition from InfoNCE to SigLIP represents a fundamental architectural shift in how multimodal alignment is learned:
InfoNCE (Softmax) uses batch-wide softmax normalization, which creates quadratic memory scaling in batch size and limits practical batch sizes to approximately 32K-64K.

It assumes mutual exclusivity within the batch and provides only implicit geometric control through temperature. Multi-objective blending is difficult because the softmax normalization interacts with additional loss terms.
SigLIP (Sigmoid) uses independent per-pair sigmoid functions, enabling linear memory scaling and batch sizes exceeding 1M. It assumes pairwise independence and provides explicit geometric control through margin and bias parameters. Multi-objective blending is natural because contributions are independent.
This dichotomy reveals that the choice between softmax and sigmoid is not merely an implementation detail but a fundamental design decision that shapes the entire training pipeline.

Dimension	**InfoNCE (Softmax Paradigm)**	**SigLIP (Sigmoid Paradigm)**
Memory Scaling	Quadratic (O(K^2) matrix materialization overhead).	Linear (treats image-text pairs independently).
Batch Constraints	Typically bounded to ~32K–64K without multi-level tiling hardware architectures.	Scalable past 1 Million samples smoothly due to independent loss terms.
Geometric Space Control	Implicitly managed via the sharpness of the temperature parameter.	Explicitly controlled via dimension-cardinality bounds and explicit margin (m, b) separation.
Sample Independence	Assumes full mutual exclusivity across the entire batch sequence.	Models pairwise independence, making it highly robust to semantic noise and looseness.

The Contrastive-Non-Contrastive Spectrum

The landscape of multimodal objectives spans a spectrum from pure contrastive to fully non-contrastive approaches:
Pure contrastive methods (CLIP, OpenCLIP) use batch negatives with InfoNCE softmax.
Hard negative enhanced methods (MoCHi, DropMix, B3) use mined or synthesized

negatives with improved weighting. Relaxed contrastive methods (ReCo) modify the energy function without abandoning the contrastive framework. Sigmoid contrastive methods (SigLIP, SigLIP 2) use pairwise independent sigmoid losses. Non-contrastive methods (NOVA) eliminate negatives entirely, using prediction and distributional regularization. Hybrid methods (SigLIP 2, MMCL) blend multiple objectives.
This spectrum reveals that the field is not converging on a single optimal objective but rather developing a toolkit of approaches suited to different constraints: batch size, domain specificity, computational budget, and downstream task requirements.

Critical Success Factors

Across the landscape of emerging objectives, several patterns emerge:

Batch independence: Methods that reduce or eliminate dependence on batch-wide statistics (sigmoid, non-contrastive) exhibit superior scaling properties.
Explicit geometric control: Losses with explicit margin parameters (SigLIP's m and b) provide more interpretable and stable training dynamics than temperature-dependent softmax approaches.
Multi-objective synergy: The most performant contemporary models blend contrastive, generative, and self-supervised objectives rather than relying on any single loss.
Domain adaptation: The optimal objective varies by domain. Medical imaging benefits from non-contrastive approaches (NOVA) where negative assumptions are violated, while general vision-language benefits from sigmoid scaling (SigLIP).

Open Challenges and Research Agenda

Theoretical Understanding

Despite empirical progress, theoretical understanding of multimodal loss landscapes remains limited:
Loss landscape geometry: How do different objectives shape the embedding manifold? What is the relationship between loss geometry and downstream task performance?
Saturation phenomena: Why do extremely large batch sizes degrade performance? Is there an optimal negative-to-positive ratio that varies by dataset or model scale?
Modality asymmetry: Vision and language encoders have fundamentally different architectures and inductive biases. How should loss design account for this asymmetry?

Scaling and Efficiency

Sub-quadratic alternatives: Can we develop objectives that scale sub-linearly with batch size while preserving the discriminative power of large negative sets?
Dynamic objective scheduling: Should the loss formulation change during training (e.g.,

starting with contrastive and transitioning to non-contrastive)?
Memory-efficient hard negative retrieval: How can global hard negative mining be made practical for trillion-scale datasets?

Semantic Fidelity

Soft negative modeling: Can we develop losses that explicitly model partial relevance rather than treating all non-pairs as equally negative?
Multi-label contrastive learning: How should objectives adapt when samples have multiple valid alignments (e.g., an image with multiple correct captions)?
Causal representation learning: Can contrastive objectives be extended to capture causal relationships between modalities rather than mere statistical associations?

Integration with Generative Objectives

The most promising frontier is the seamless integration of contrastive and generative objectives:
Diffusion-based alignment: Can diffusion models serve as both generative and discriminative objectives, unifying generation and representation learning?
Autoregressive multimodal modeling: How should next-token prediction in language models be combined with contrastive alignment across modalities?
Reinforcement learning from human feedback: Can RLHF objectives for multimodal models be formulated as generalized contrastive learning?

Towards Adaptive Objective Functions

As multimodal models continue scaling to trillions of parameters, objective functions themselves are becoming increasingly important architectural components rather than simple optimization tools. Future loss functions are likely to combine contrastive alignment, probabilistic embedding regularization, uncertainty estimation, curriculum learning, reinforcement learning, and self-distillation within unified optimization frameworks. Instead of selecting a single objective throughout training, foundation models may automatically learn which objectives to optimize based on model maturity, dataset characteristics, computational resources, and downstream deployment requirements. The evolution from InfoNCE to adaptive objective learning represents not merely an incremental improvement in optimization, but the emergence of a new paradigm in which the loss function itself becomes an intelligent, dynamically evolving component of the learning process.

Conclusion

The evolution from InfoNCE to next-generation objectives for multimodal LLMs represents a microcosm of broader trends in AI: the pursuit of scale, the demand for efficiency, and the quest for theoretical understanding. What began as a simple softmax-based contrastive loss has ramified into a rich landscape of sigmoid-based, non-contrastive, and hybrid approaches - each addressing specific limitations of the original formulation while introducing new trade-offs.

The field is converging toward several key principles: batch-independent normalization, explicit geometric control, multi-objective synergy, and domain-adaptive design. SigLIP's sigmoid approach has demonstrated production viability at unprecedented scale. NOVA's non-contrastive framework offers simplicity and stability in domains where negative assumptions fail. LLM2CLIP shows how the capabilities of large language models can be distilled into embedding spaces. And hybrid approaches like SigLIP 2 point toward a future where discriminative and generative objectives are unified.

For practitioners, the recommendation is: evaluate sigmoid-based losses for new vision-language projects, consider non-contrastive alternatives for specialized domains with noisy alignments, and embrace multi-objective training for state-of-the-art results. For researchers, the open challenges in theoretical understanding, scaling efficiency, and semantic fidelity define the frontier.

The transition from InfoNCE to beyond is not merely an engineering optimization but a paradigm shift in how we conceptualize multimodal learning. The next generation of multimodal foundation models are likely to be trained with objectives that bear little resemblance to the simple contrastive loss that launched the field. Yet the fundamental insight, that alignment across modalities can be learned through comparative supervision, remains as powerful as ever.

References

Cheng, Z., Zhang, H., Li, K., Leng, S., Hu, Z., Wu, F., ... & Bing, L. (2024). Breaking the Memory Barrier: Near Infinite Batch Size Scaling for Contrastive Loss. *arXiv preprint arXiv:2410.17243*.

Wang, T., & Isola, P. (2020, November). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning* (pp. 9929-9939). PMLR.

Yang, S., Du, J., Lu, S., Zhang, W., Wang, N., & Li, H. (2025). CLIPin: A Non-contrastive Plug-in to CLIP for Multimodal Semantic Alignment. *arXiv preprint arXiv:2508.06434*.

Bose, A. J., Ling, H., & Cao, Y. (2018, July). Adversarial contrastive estimation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1021-1032).

Kalantidis, Y., Sariyildiz, M. B., Pion, N., Weinzaepfel, P., & Larlus, D. (2020). Hard negative mixing for contrastive learning. *Advances in neural information processing systems*, *33*, 21798-21809.

Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. (2023). Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 11975-11986).

Kuhn, L., Serra, G., & Buettner, F. (2026). Non-Contrastive Vision-Language Learning with Predictive Embedding Alignment. *arXiv preprint arXiv:2602.00653*.

Lin, Z., Bas, E., Singh, K. Y., Swaminathan, G., & Bhotika, R. (2023). Relaxing contrastiveness in multimodal representation learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2227-2236).

Huang, W., Wu, A., Yang, Y., Luo, X., Yang, Y., Hu, L., ... & Qiu, L. (2024). Llm2clip: Powerful language model unlocks richer visual representation. *arXiv preprint arXiv:2411.04997*.