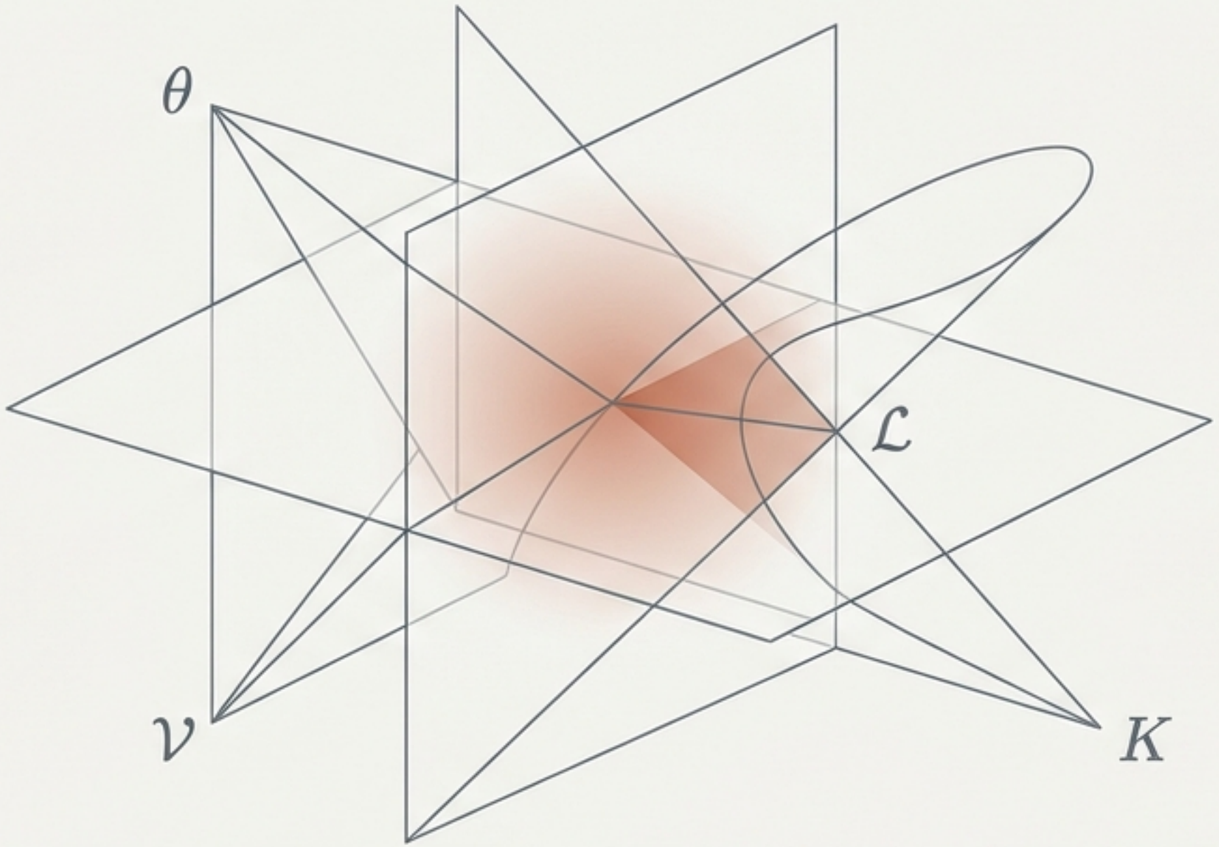
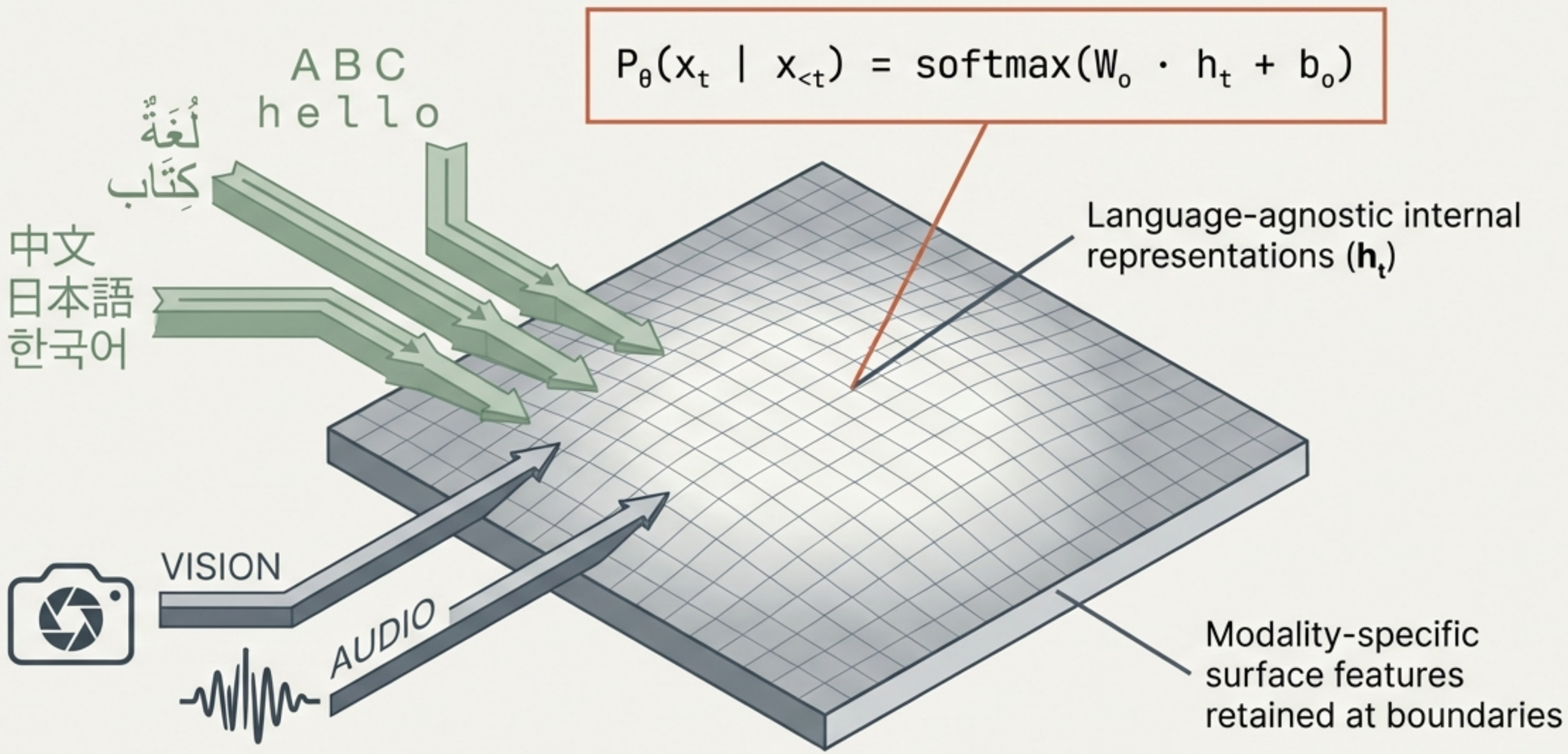
The Multimodal and Multilingual Frontier

Architecture, Alignment, and the Future of Generalist LLMs



THE SHARED REPRESENTATIONAL MANIFOLD



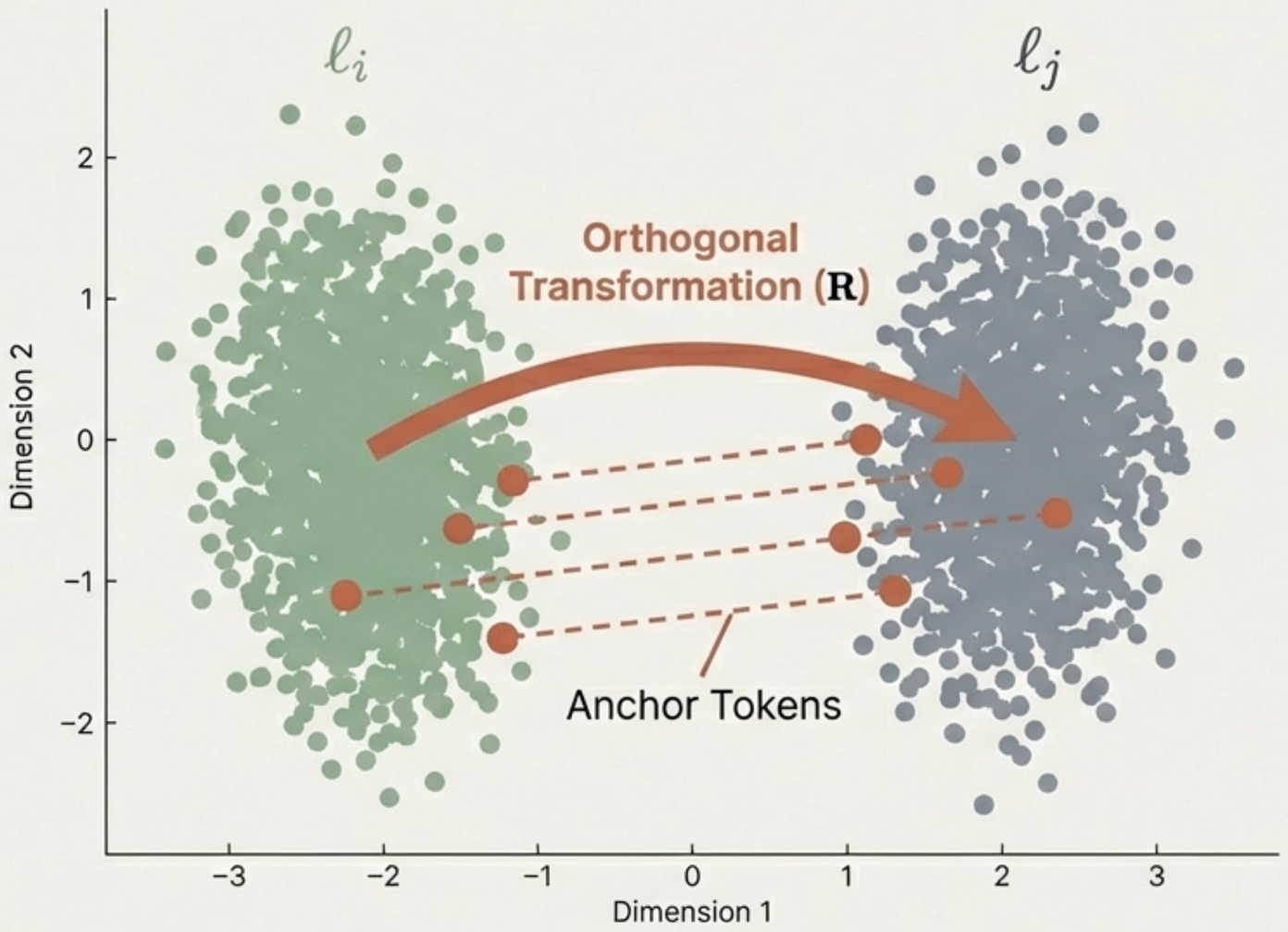
MULTILINGUAL GEOMETRY & ZERO-SHOT TRANSFER

Procrustes alignment condition:

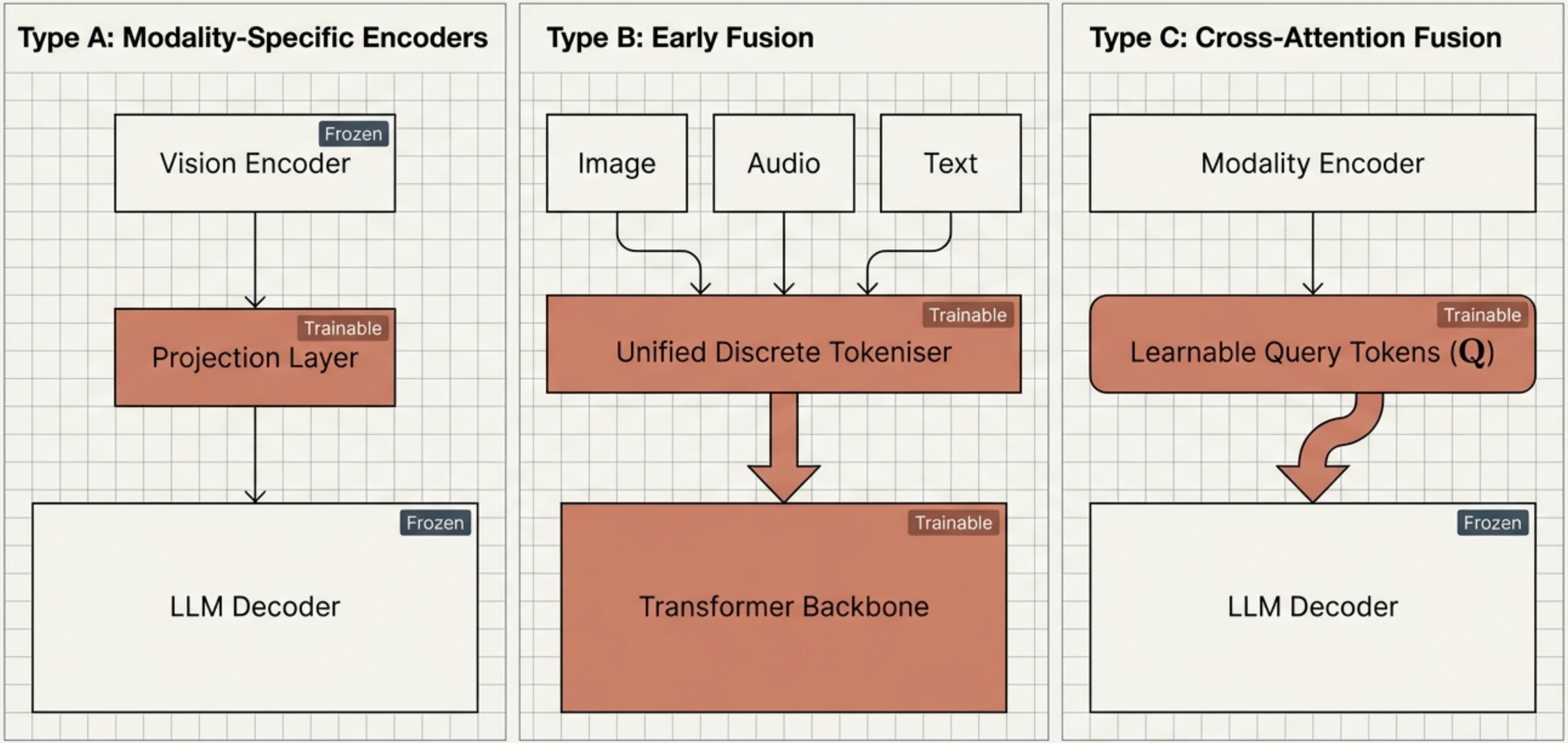
$$\left\|\mathbf{H}^{\ell_i}\mathbf{R} - \mathbf{H}^{\ell_j}\right\|_F^2 < \epsilon$$

Key Insight:

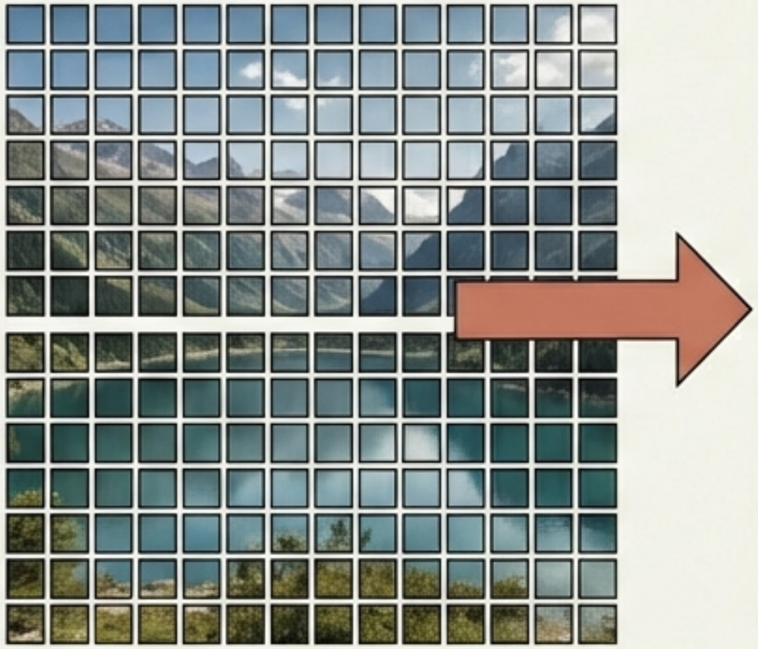
- Knowledge transfers without target-language supervision because shared vocabulary (“Anchor Tokens”) forces local alignment across the latent space.

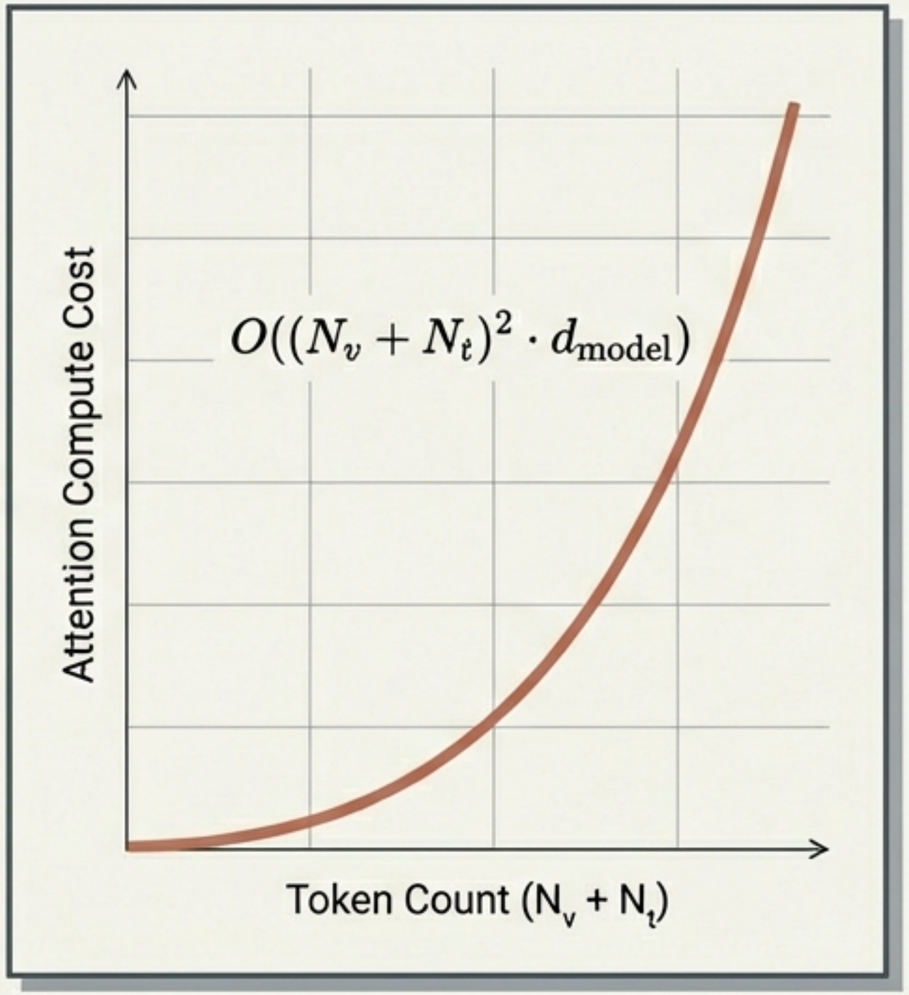


MULTIMODAL ARCHITECTURAL TAXONOMY



THE INFORMATION BOTTLENECK & VISUAL TOKEN BUDGETS





$$N_v = \frac{H \times W}{p^2}$$

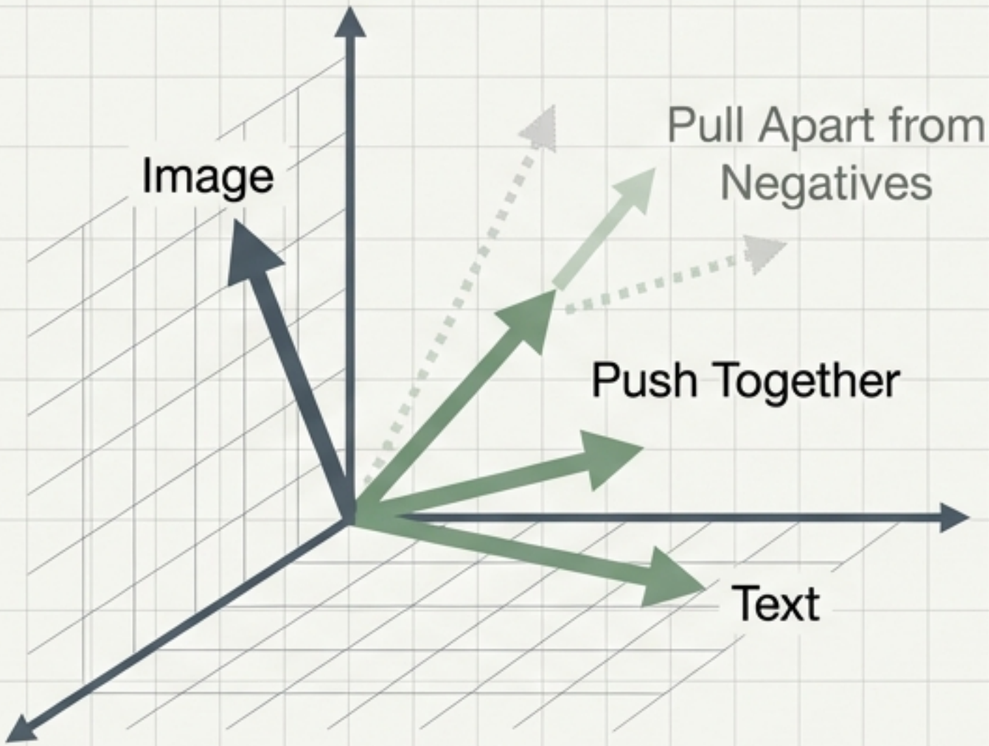
e.g., 336×336 ViT-L/14 yields **576 tokens**

Compression Strategies:

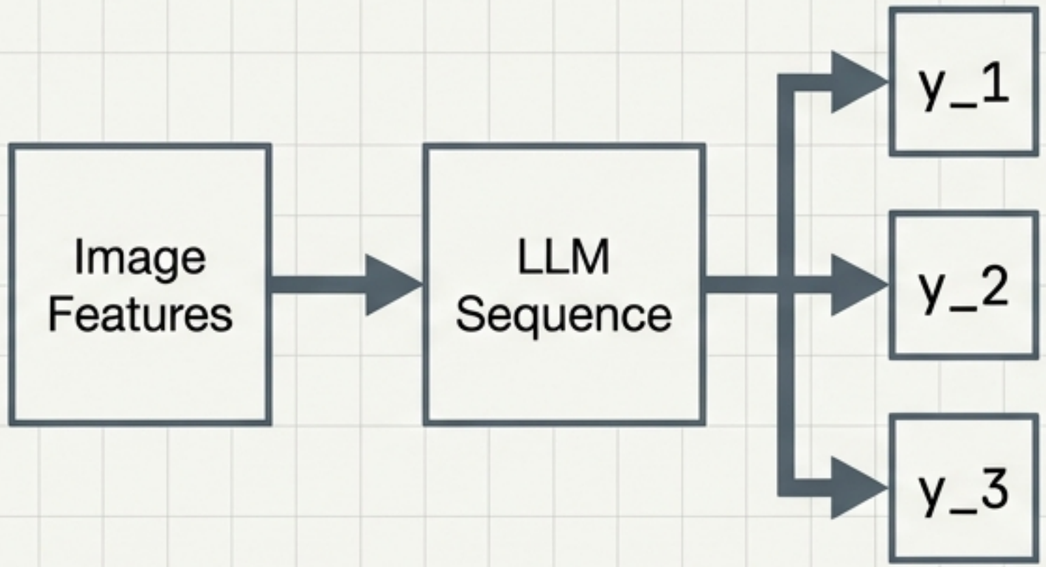
1. Perceiver Resampler (Fixed N_q)
2. Average Pooling (Reduced N_v')
3. Adaptive Tokenisation
4. Tile-and-Merge

MECHANISMS OF CROSS-MODAL ALIGNMENT

Contrastive Alignment ($\mathcal{L}_{CLIP}$)



Generative Alignment ($\mathcal{L}_{gen}$)



THE TENSION:
Maximising cross-modal alignment forces modality-invariant features, while generative fidelity requires modality-specific detail.

THE PROGRESSIVE TRAINING PIPELINE

STAGE 0:
Unimodal Pre-training
(CLM)

STAGE 1:
Modality Alignment
(Caption generation, ~1e-3 LR)

STAGE 2:
Instruction Tuning
(~2e-5 LR)

STAGE 3:
Preference Optimisation
(DPO/RLHF, ~5e-7 LR)

Critical Warning: Direct joint training causes catastrophic representation collapse.

Data Mixing Strategy:

$$\lambda_k \propto \frac{\texttt{task importance}_k}{\sqrt{\texttt{dataset size}_k}}$$

Inverse-square-root scaling prevents massive datasets from dominating.

THE CURSE OF MULTILINGUALITY & TOKENISATION INEQUITY

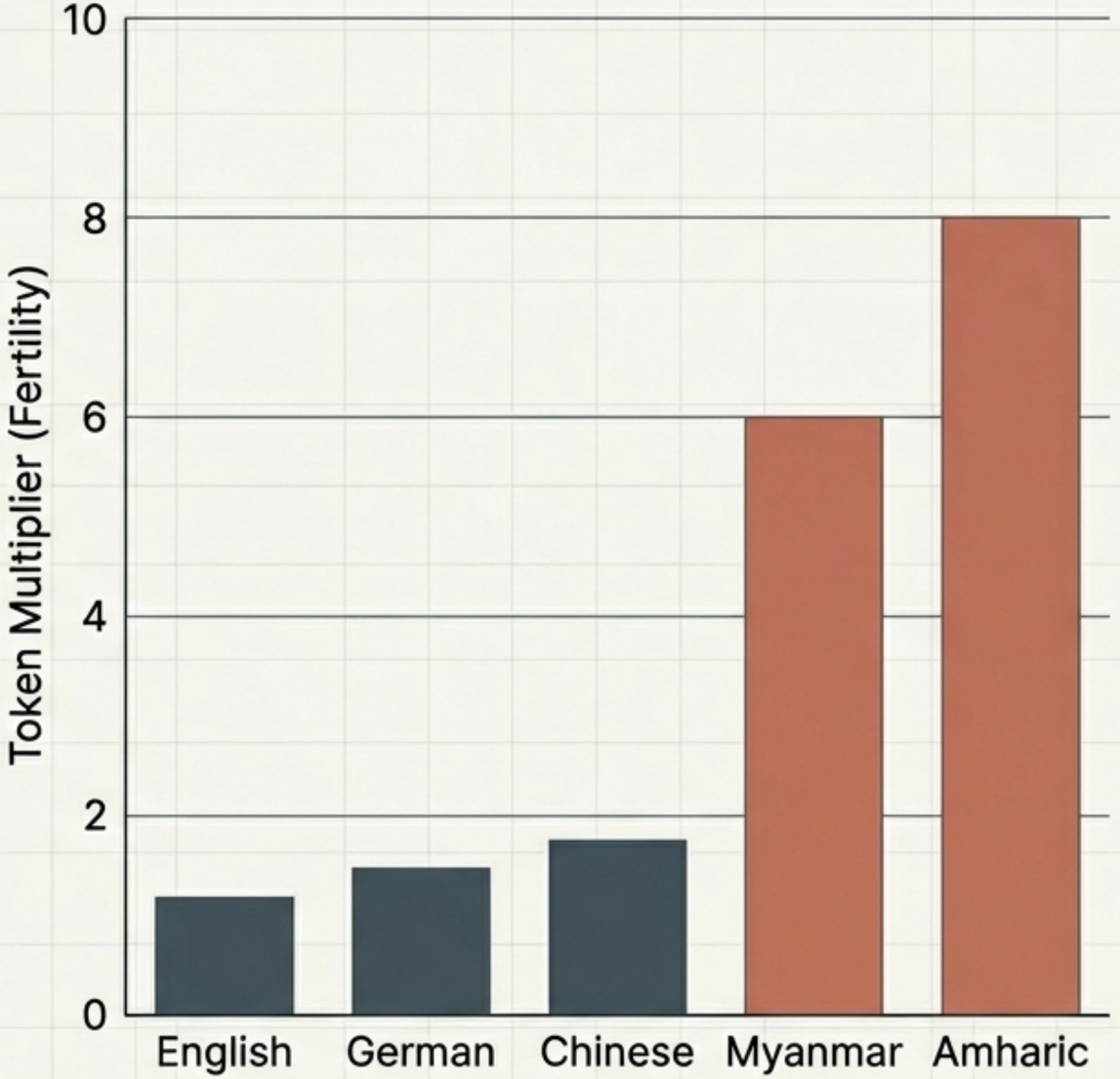
THE CAPACITY TRADE-OFF

Adding more languages (L) degrades high-resource performance once critical capacity $(\rho < \rho_{crit})$ is breached.

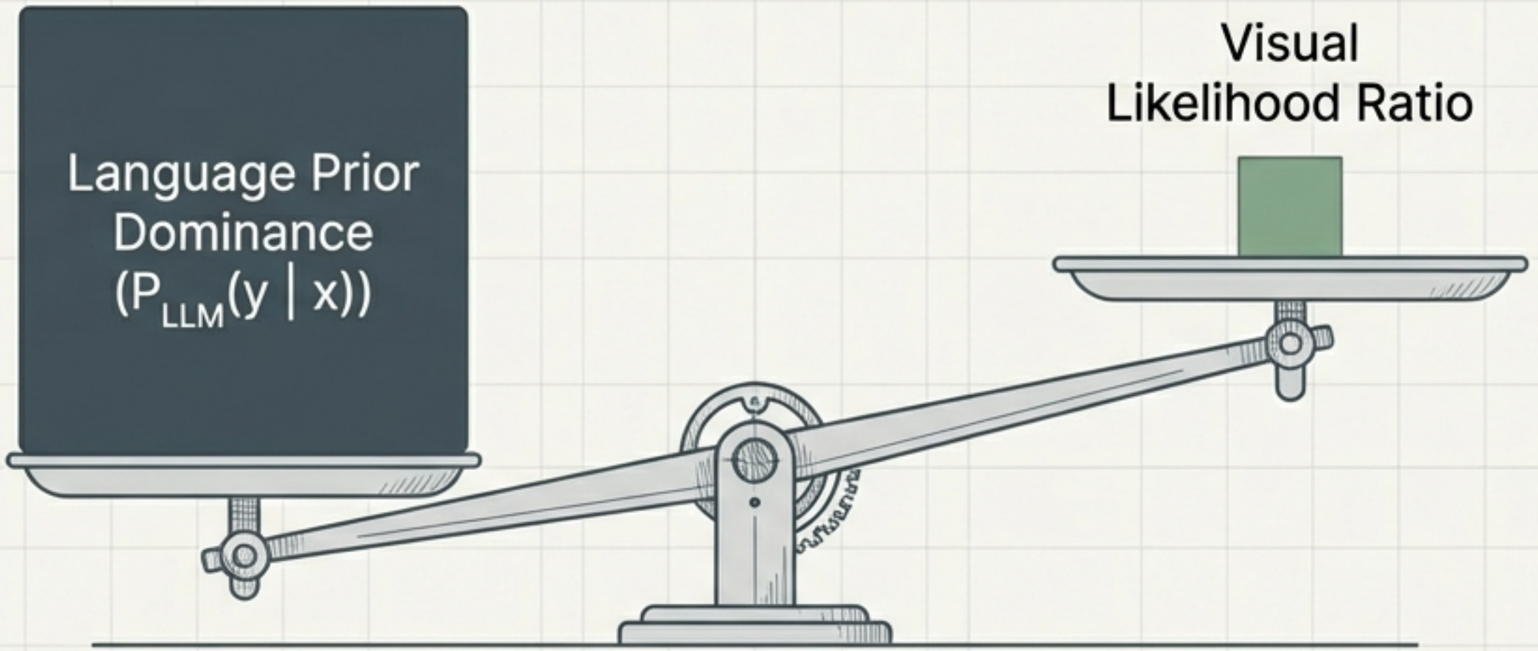
IMPACT

High-fertility languages suffer rapid context window exhaustion and drastically inflated inference costs.

TOKEN MULTIPLIER (FERTILITY)



Modality Gaps & Object Hallucination



Core Vulnerability: When visual evidence is weak, linguistically plausible priors override reality, generating objects not present in the input.

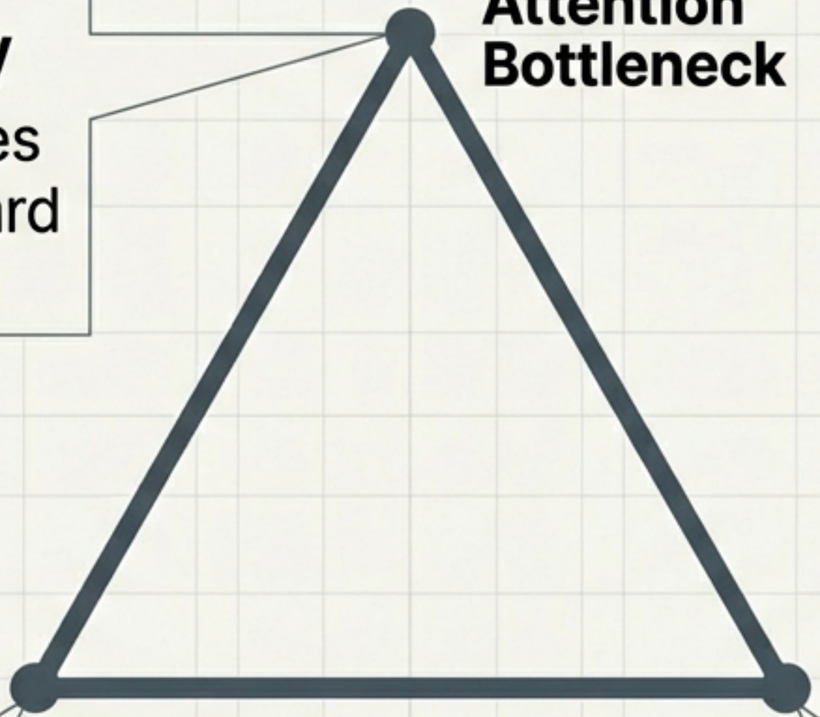
Evaluation Metric: Evaluated via CHAIR (Caption Hallucination Assessment with Image Relevance), measuring hallucinated objects vs. total objects.

Architectural Limits at Scale

$O((N_v + N_t)^2)$ Complexity

32 frames of video generates ~18K tokens, making standard attention intractable.

Attention Bottleneck



Autoregressive Limits

Flattened Awareness

Spatial Blindness

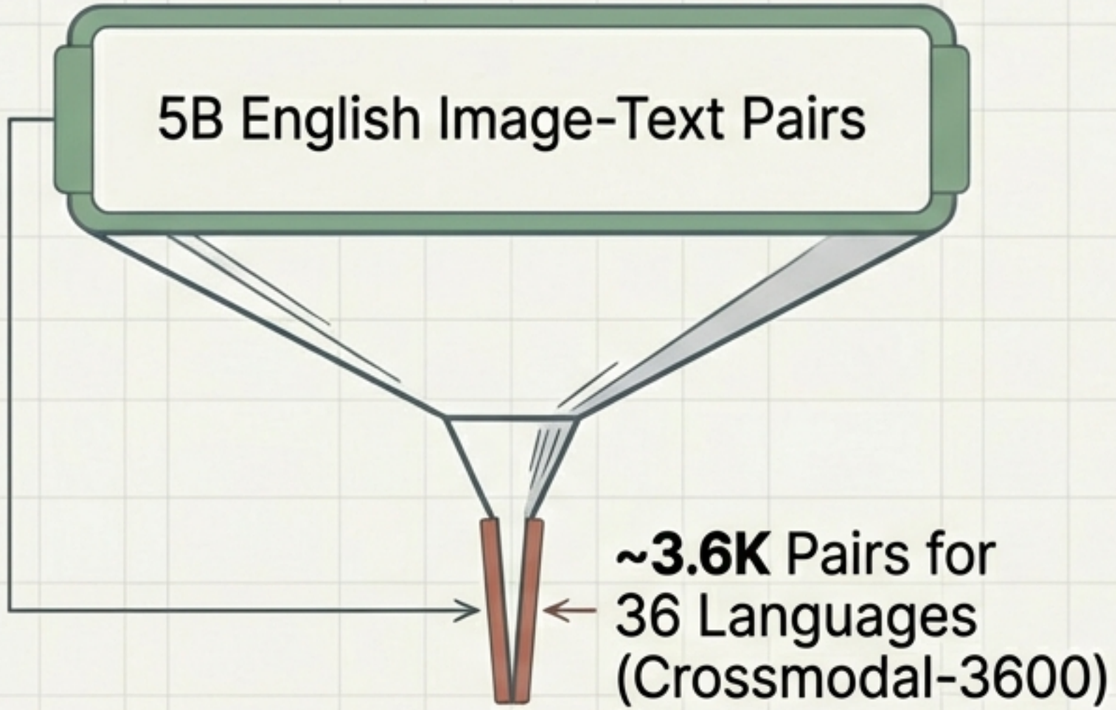
Transformers flatten 2D inputs into 1D sequences, forcing implicit—rather than native—spatial reasoning.

Sequential Bottleneck

Generating a 1024x1024 image autoregsively demands ~16K sequential steps.

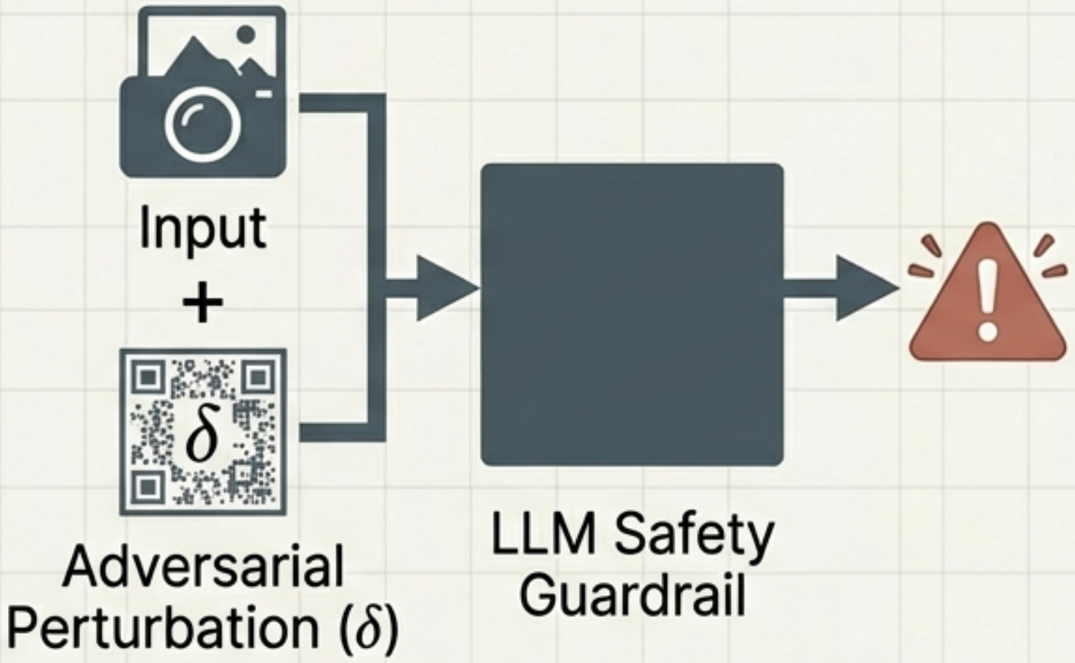
Safety Vulnerabilities & Data Scarcity

Compounding Scarcity:



Low-resource languages lack both text volumes and multimodal paired data.

Cross-Modal Jailbreaking:

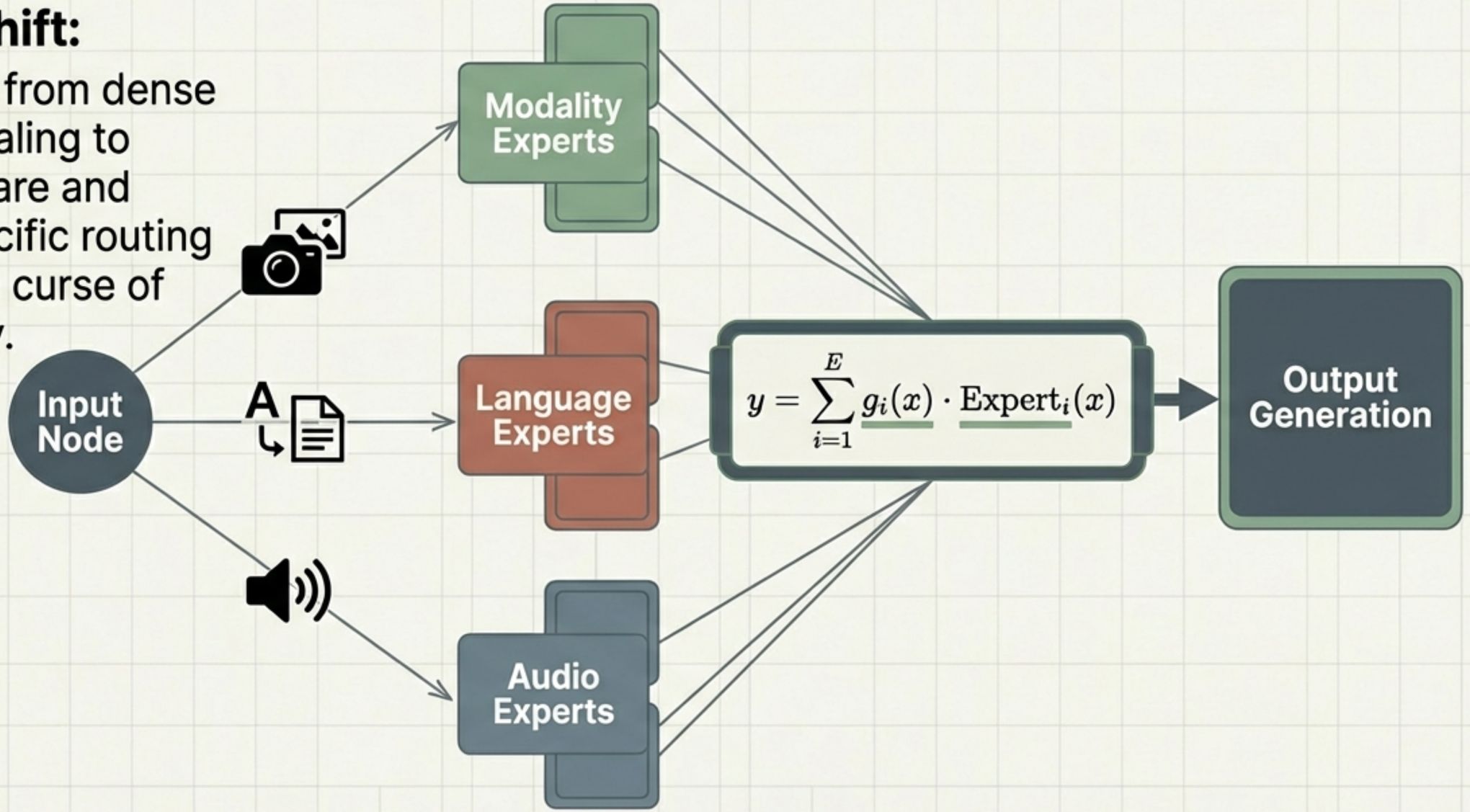


Multimodal inputs introduce severe attack surfaces, while safety alignment remains overwhelmingly English-centric.

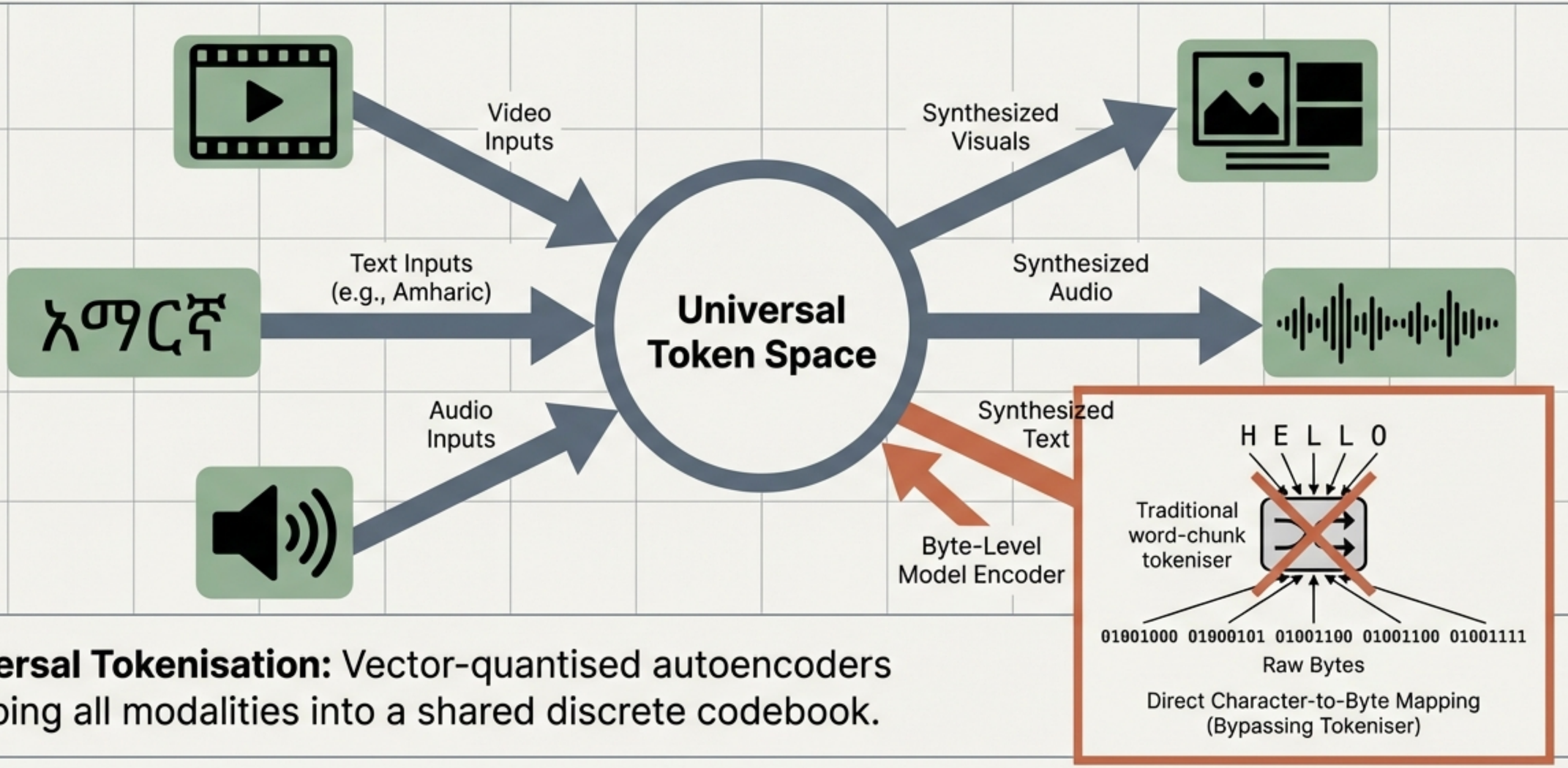
Frontier: Sparse Routing & Mixture-of-Experts

Strategic Shift:

Moving away from dense parameter scaling to language-aware and modality-specific routing to bypass the curse of multilinguality.



Unified Any-to-Any & Byte-Level Architectures



Universal Tokenisation: Vector-quantised autoencoders mapping all modalities into a shared discrete codebook.

Eliminating Inequity: Byte-level models (e.g., ByT5) achieve $\mathrm{Fertility}(\ell) = 1$ **across all languages, completely** solving English tokenisation bias.

Agentic Systems & Neuro-Symbolic Reasoning

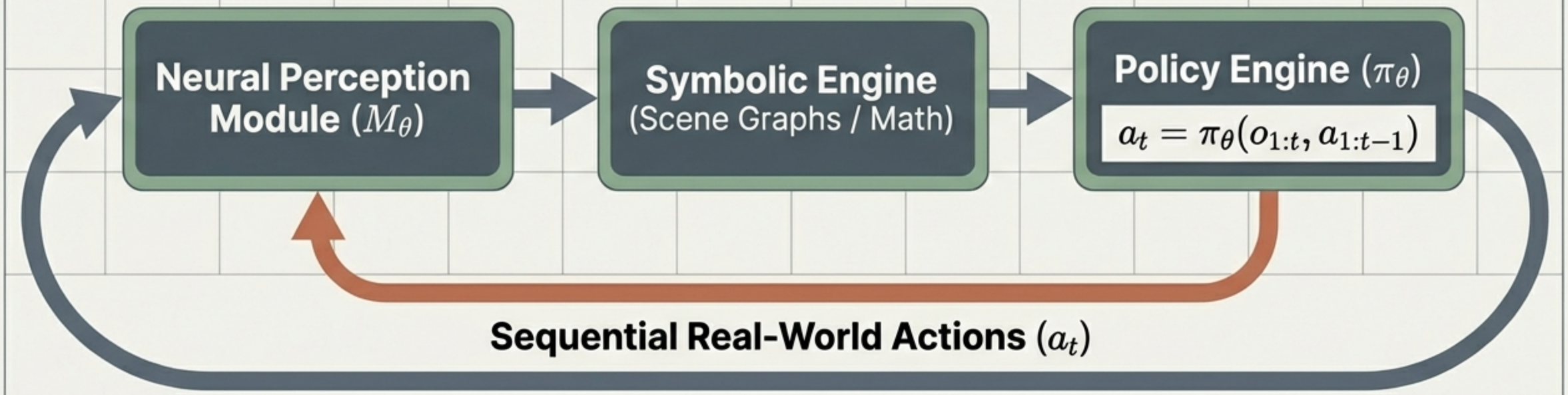
Embodied AI:

Models calculating real-time actions from live visual/audio streams:

$$a_t = \pi_\theta(o_{1:t}, a_{1:t-1})$$

Symbolic Grounding:

Integrating LLMs with structured reasoning engines to execute formal spatial and mathematical logic natively.



Strategic Summary Matrix

Category	Core Limitation	Root Cause	Frontier Mitigation
Representational	Language collapse	Fixed capacity	**MoE & Adapters**
Tokenisation	Fertility inequity	BPE trained on English	**Byte-level models**
Architecture	Sequential bottlenecks	Attention complexity	**Sparse attention & Parallel decoding**
Safety	Cross-modal jailbreaks	English-centric alignment	**Multilingual red-teaming**