VERACE PRIVATE LIMITED

VERACE V1: MULTIMODAL LANGUAGE MODEL WITH SPECTRAL DIFFERENTIAL ATTENTION, HOLOGRAPHIC MEMORY, MANIFOLD EXPERTS, AND ADAPTIVE COGNITIVE DEPTH

Krrish Choudhary
Chief Executive Officer
Verace Private Limited
krrish@verace.in

August 3, 2026

ABSTRACT

Standard Transformer architectures face severe physical and computational bottlenecks at scale: quadratic memory expansion of key-value (KV) caches, attention matrix degradation over long horizons, massive all-to-all communication overheads in discrete Mixture-of-Experts (MoE) routing, and static per-token compute allocation regardless of token complexity. In this paper, we present **Verace V1**, a next-generation multimodal large language model architecture engineered top-down to resolve these architectural limits. Verace V1 replaces standard softmax self-attention with **Spectral State-Space Differential Attention (SSSD)**, which maintains a norm-conserving recurrent state via a skew-symmetric Cayley transform ($||\Psi_t||_F = ||\Psi_0||_F$). It eliminates KV-cache expansion using **Continuous Holographic Associative Memory (CHAM)**, an $O(1)$-memory complex matrix on the unitary Lie group $U(d)$ updated via Newton-Schulz polynomial retraction ($H^H H = I$). Distributed expert routing bottlenecks are removed by **Manifold Continuous MoE (M-CMoE)**, which generates per-token low-rank expert weight adaptations locally from a shared manifold basis ($U_j \text{diag}(\sigma_j(x)) V_j^T$) without cross-device dispatch. To dynamicize compute allocation, the **Adaptive Cognitive Depth Engine (ACDE)** implements per-token early exit (2 to 128 layers) via per-example token index gathering, reducing FLOPs while guaranteeing mathematical batch independence. Generative decoding and pretraining are guided by a scalar **Latent Energy Critic** ($E(x, y) = ||h_{\text{cand}} - W_{\text{energy}} x||_2^2$) executing parallel latent tree search. Parameter optimization is governed by the **Unitary Muon Optimizer**, which projects 2D weight updates onto the Stiefel manifold via exact SVD polar decomposition, overcoming the permanent non-convergence oscillations of iterative polynomial methods on rectangular weight matrices. We formally detail the architectural derivations, prove exact algebraic invariants, describe Triton GPU acceleration paths, provide complete pseudocode, and validate empirical stability across our complete unit test suite.

Keywords Multimodal Architectures · State-Space Models · Holographic Memory · Continuous Mixture of Experts · Adaptive Compute · Manifold Optimization · Energy-Based Models

1 Introduction

The dominant paradigm in artificial intelligence relies heavily on the standard Transformer architecture [1]. While scaling laws have demonstrated remarkable performance gains across language and multimodal reasoning tasks [2], standard Transformers suffer from fundamental structural limitations rooted in their core mathematical formulations:

1. **Quadratic Attention & KV-Cache Memory Growth:** Softmax self-attention exhibits $O(T^2)$ time complexity for sequence length T. At inference time, storing key-value pairs (the KV cache) requires memory that scales linearly with sequence length and batch size ($O(B \cdot T \cdot d_{\text{kv}})$). For ultra-long context windows, KV-cache memory bandwidth becomes the primary bottleneck, exceeding GPU VRAM capacity.
2. **State Norm Instability in Linear Attention:** Linear attention and state-space models (SSMs) [3] compress context into a fixed-size recurrent matrix to achieve $O(1)$ memory complexity. However, unconstrained recurrent update matrices suffer from vanishing or exploding state norms over long context horizons, causing numerical drift and loss of long-range recall.
3. **All-to-All Dispatch Latency in Sparse MoE:** Mixture-of-Experts architectures [4] scale model capacity without proportional FLOP increases by routing tokens to discrete sub-networks. In distributed clusters, routing tokens across devices demands expensive all-to-all collective communication, severely degrading hardware efficiency and throughput.
4. **Static Per-Token Compute Allocation:** Standard Transformers process every token through the exact same number of layers N. Simple tokens (e.g., punctuation or repeated prepositions) consume identical compute as highly complex logical reasoning tokens, causing severe computational waste.
5. **Autoregressive Greedy Decoding Drift:** Standard token-by-token sampling commits early to sampling choices without evaluating long-term consistency, causing hallucinations and error accumulation.

To overcome these structural limits, we introduce **Verace V1**, an architecture designed top-down: beginning with exact targeted properties (such as strictly invariant state norms, $O(1)$ memory footprints, local continuous expert generation, and deterministic batch independence), and constructing mathematical mechanisms that provably satisfy them.

The reference implementation and complete code resources are open-sourced at `https://github.com/Verace-Pvt-Ltd/Verace-V1`, with the architectural paper hosted at `https://verace.in/research/verace-v1-architecture`.

—

2 System Architecture & Component Flow Diagrams

The complete system topology of Verace V1 is illustrated in Figure 1. The model processes input text and optional visual inputs through dedicated tokenizers/encoders, fuses visual representations non-destructively, passes sequence representations through an adaptive stack of N decoder layers governed by ACDE, and outputs next-token logits alongside Latent Energy Critic guidance.

Table 1: Comparison of Standard Transformer Subsystems vs. Verace V1 Innovations.

Subsystem	Standard Transformer	Verace V1 Architecture
Sequence Mixing	Softmax Self-Attention ($O(T^2)$)	Spectral State-Space Differential Attention (SSSD)
Long-Range Memory	KV Cache ($O(T)$ Memory Growth)	Continuous Holographic Associative Memory (CHAM, $O(1)$)
Expert Computation	Discrete Routed MoE (All-to-All)	Manifold Continuous MoE (M-CMoE, Local Basis)
Compute Allocation	Fixed Per-Layer Depth (N layers)	Adaptive Cognitive Depth Engine (ACDE, Early Exit)
Decoding Strategy	Greedy / Top-p / Beam Search	Latent Energy Critic (Latent Tree Search)
Vision Processing	Separate ViT / Linear Projection	Patch ViT + 2×2 Token Merge + Non-Destructive Fusion
Optimization	AdamW / Standard Muon	Unitary Muon (Exact SVD Polar Stiefel Projection)

—

3 Spectral State-Space Differential Attention (SSSD)

To eliminate the $O(T^2)$ self-attention bottleneck without incurring numerical drift, SSSD attention maintains a per-head state matrix $\Psi_t \in \mathbb{R}^{d_h \times d_h}$ updated at every step via an **exact orthogonal rotation** $R_t \in SO(d_h)$, building upon linear state-space sequence models [3] (as illustrated in Figure 2).

Given input representation $x_t \in \mathbb{R}^{d_{\text{model}}}$, linear projections yield query q_t, key k_t, value $v_t \in \mathbb{R}^{d_h}$, and continuous rate parameters $\delta_t \in (0, 0.1]$:

$$q_t = W_q x_t, \quad k_t = W_k x_t, \quad v_t = W_v x_t, \quad \delta_t = 0.1 \cdot \sigma(W_\delta x_t) \tag{1}$$

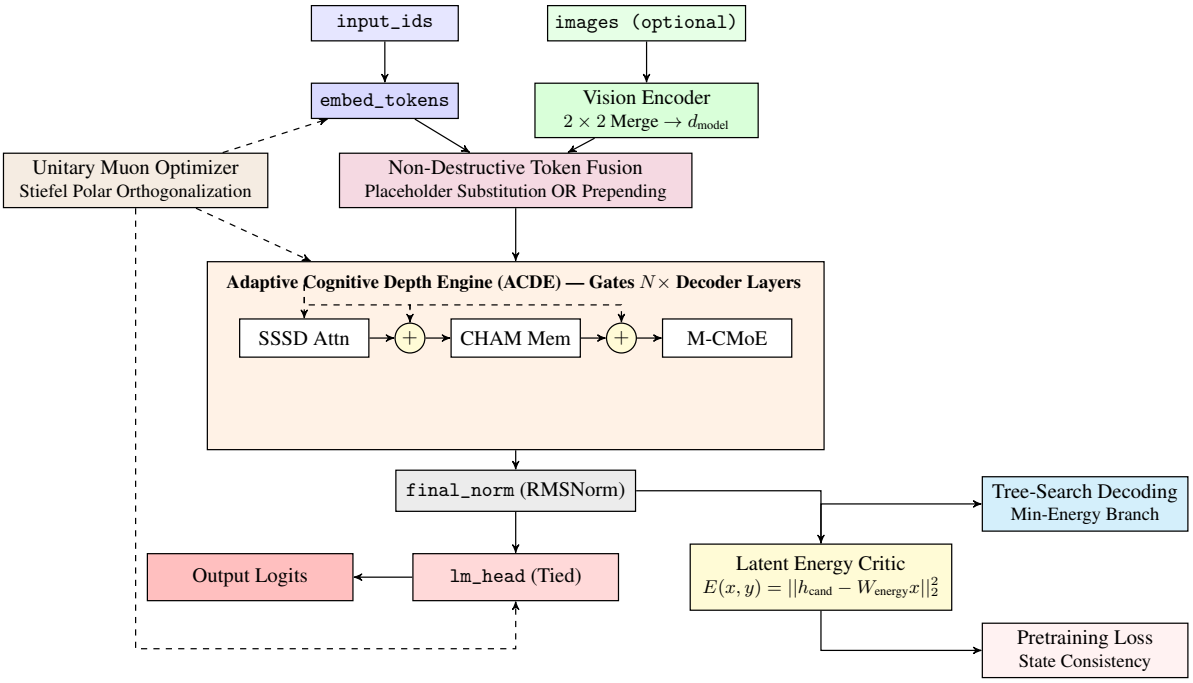

Figure 1: Full end-to-end system architecture of Verace V1 demonstrating input processing, multimodal token fusion, ACDE dynamic decoder stack, energy-guided inference/training, and Unitary Muon parameter optimization.

Table 2: Verace V1 Model Architecture & Hyperparameter Specifications.

Hyperparameter Field	**Architectural Function**
`vocab_size`	Vocabulary dimension for token embedding and LM head projection
`hidden_dim`	Model hidden state width (d_{model}) across layers
`num_layers`	Maximum instantiated decoder layer depth (N_{max})
`num_heads` / `head_dim`	SSSD attention heads (H_{sssd}) and per-head state dimension (d_{head})
`chams_holographic_dim`	Dimension of complex matrix hologram $H \in U(d_{\text{holo}})$
`mcmoe_rank` / `mcmoe_num_components`	Low-rank basis generator rank (r) and component bank size (N)
`top_k_components`	Number of active basis components selected per token (K)
`min_cognitive_depth` / `max_cognitive_depth`	Lower (N_{min}) and upper (N_{max}) per-token layer execution bounds
`energy_halting_threshold`	Halting probability remainder threshold (τ_{halt})
`tree_branches`	Candidate branches scored per step during Latent Tree Search
`depth_penalty_weight` / `energy_penalty_weight`	Pretraining multi-objective loss scaling coefficients (λ_{depth}, λ_{energy})
`learning_rate` / `unitary_muon_momentum`	Unitary Muon optimizer learning rate (η) and momentum (μ)
`weight_decay` / `rms_norm_eps`	Decoupled weight decay (λ_{wd}) and RMSNorm stability epsilon (ϵ_{norm})

From key k_t and value v_t, we construct a rank-2 skew-symmetric generator matrix $A_t \in \mathbb{R}^{d_h \times d_h}$:

$$A_t = k_t v_t^T - v_t k_t^T \implies A_t^T = -A_t \tag{2}$$

The state rotation R_t is obtained via the **Cayley transform** of A_t:

$$R_t = \left(I - \frac{\delta_t}{2} A_t\right)^{-1} \left(I + \frac{\delta_t}{2} A_t\right) \tag{3}$$

The state matrix Ψ_t and output o_t evolve according to:

$$\Psi_t = R_t \Psi_{t-1}, \qquad o_t = \Psi_t q_t \tag{4}$$

Theorem 1 (Exact State Norm Conservation)**.** *For any skew-symmetric matrix $A_t^T = -A_t$ and $\delta_t \in \mathbb{R}$, the Cayley transform R_t is strictly orthogonal ($R_t^T R_t = I$). Consequently, the Frobenius norm of Ψ_t is strictly invariant for all $t \geq 0$:*

$$||\Psi_t||_F = ||\Psi_0||_F \tag{5}$$

Proof. Consider $R_t^T R_t$:

$$R_t^T = \left(I + \frac{\delta_t}{2}A_t\right)^T \left(I - \frac{\delta_t}{2}A_t\right)^{-T} = \left(I - \frac{\delta_t}{2}A_t\right)\left(I + \frac{\delta_t}{2}A_t\right)^{-1} \tag{6}$$

$$R_t^T R_t = \left(I - \frac{\delta_t}{2}A_t\right)\left(I + \frac{\delta_t}{2}A_t\right)^{-1}\left(I - \frac{\delta_t}{2}A_t\right)^{-1}\left(I + \frac{\delta_t}{2}A_t\right) = I \tag{7}$$

Since $R_t \in SO(d_h)$ is an isometry, $||\Psi_t||_F^2 = \text{Tr}(\Psi_t^T \Psi_t) = \text{Tr}(\Psi_{t-1}^T R_t^T R_t \Psi_{t-1}) = \text{Tr}(\Psi_{t-1}^T \Psi_{t-1}) = ||\Psi_{t-1}||_F^2$. By induction, $||\Psi_t||_F = ||\Psi_0||_F$. □

When an active mask position is false (halted token), δ_t is set to zero, forcing $R_t = I$, so the state recurrence seamlessly becomes an identity update.

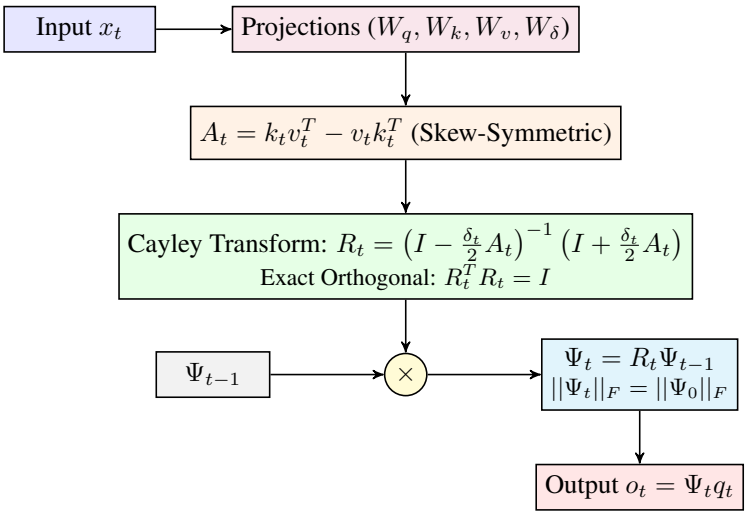

Figure 2: SSSD Attention per-timestep state rotation via skew-symmetric Cayley transformation.

—

4 Continuous Holographic Associative Memory (CHAM)

To replace the growing KV cache of standard self-attention [1] with an $O(1)$ memory footprint, CHAM maintains a single complex matrix $H_t \in \mathbb{C}^{d_{\text{holo}} \times d_{\text{holo}}}$ constrained to the unitary group $U(d_{\text{holo}})$, satisfying $H_t^H H_t = I$ (see Figure 3).

1. **Infinitesimal Outer-Product Write:** Given projections $k_t, v_t \in \mathbb{R}^{d_{\text{holo}}}$ and continuous gating parameter $\gamma_t = 0.1 \cdot \sigma(W_\gamma x_t)$:

$$\tilde{H}_t = H_{t-1} \cdot \left(I + i \cdot \gamma_t (k_t v_t^T)\right) \tag{8}$$

2. **Newton-Schulz Unitary Retraction:** To prevent numerical divergence off the Lie group $U(d)$, $\tilde{H}_t$ is projected back onto $U(d)$ via k iterations of Newton-Schulz polynomial retraction ($k = 3$ default):

$$H^{(0)} = \tilde{H}_t, \qquad H^{(m+1)} = \frac{1}{2} H^{(m)} \left(3I - (H^{(m)})^H H^{(m)}\right) \tag{9}$$

This iteration exhibits quadratic convergence to the nearest unitary matrix on $U(d)$.

3. **Read Operation:** The output representation is retrieved by projecting query q_t through the real component of the updated hologram:

$$o_t = \text{Re}(H_t q_t) \tag{10}$$

Because H_t retains fixed dimensions $d_{\text{holo}} \times d_{\text{holo}}$ across arbitrary sequence lengths, CHAM achieves strict $O(1)$ spatial and temporal per-token computational complexity.

—

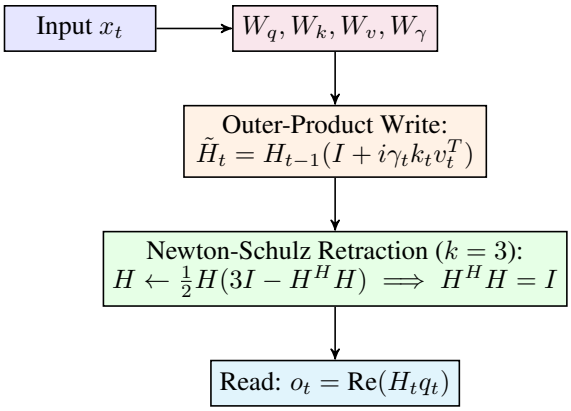

Figure 3: CHAM Memory rank-1 outer product write and Newton-Schulz unitary retraction loop.

5 Manifold Continuous Mixture-of-Experts (M-CMoE)

Conventional MoE architectures [4] suffer from distributed all-to-all network bottlenecks. M-CMoE eliminates token dispatch by locally assembling per-token low-rank weight adaptations from a shared, device-resident basis matrix bank $\{U_j, V_j\}_{j=1}^{N}$, where $U_j \in \mathbb{R}^{d_{\text{model}} \times r}$ and $V_j \in \mathbb{R}^{d_{\text{model}} \times r}$, as depicted in Figure 4.

For token vector x, a router network computes top-K component weights $\phi_j(x)$:

$$\phi(x) = \text{TopK}\left(\text{softmax}(W_{\text{gate}}x), K\right) \tag{11}$$

Concurrently, a small hypernetwork conditions on x to generate continuous singular-value scaling factors $\sigma_j(x) \in \mathbb{R}^r$:

$$\sigma_j(x) = \text{Softplus}\left(W_{\text{hyper},j}x\right) \tag{12}$$

The continuous adaptation matrix $\Delta W(x)$ is constructed locally without inter-device token movement:

$$\Delta W(x) = \sum_{j \in \text{TopK}(x)} \phi_j(x) \cdot \left(U_j \text{diag}(\sigma_j(x)) V_j^T\right) \tag{13}$$

The total expert contribution combines a base Feed-Forward Network using SiTU-GLU activations with $\Delta W(x)x$:

$$y = \text{SiTU-GLU}(xW_{\text{base}}) + \Delta W(x)x \tag{14}$$

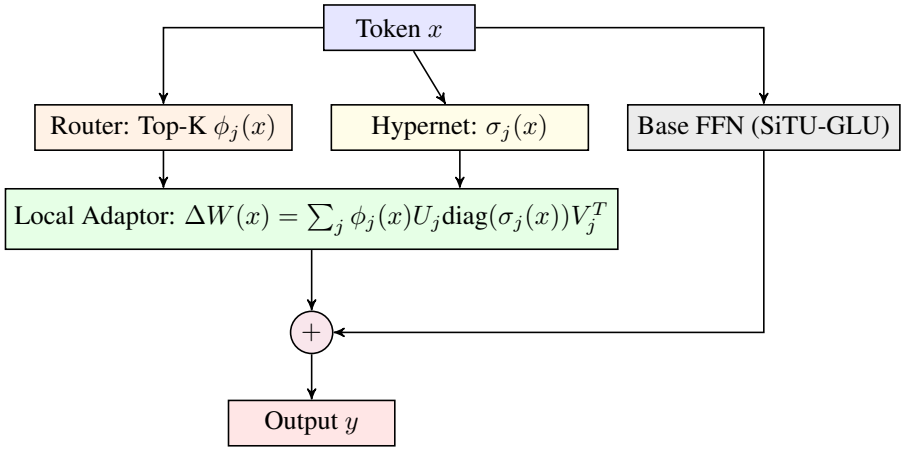

Figure 4: M-CMoE per-token continuous low-rank adaptation generated locally from a shared manifold basis.

—

6 Adaptive Cognitive Depth Engine (ACDE)

ACDE provides per-token early exit across N decoder layers (ranging between $N_{\text{min}} = 2$ and $N_{\text{max}} = 128$), extending adaptive computation time concepts [5] to grant dynamic compute allocation based on token difficulty (illustrated in Figure 5).

At each layer l, halting probability $p_{l,i}$ for active token i is calculated via:

$$p_{l,i} = \begin{cases} 0 & \text{if } l < N_{\text{min}} \\ \sigma(W_h h_{l,i}) & \text{if } l \geq N_{\text{min}} \end{cases} \tag{15}$$

Halting probabilities accumulate: $c_{l,i} = c_{l-1,i} + p_{l,i}$. When $c_{l,i} \geq 1 - \epsilon$ (where $\epsilon = 0.01$), token i halts and exits execution.

Algorithm 1 ACDE Per-Example Gathering and Execution Loop

Input: Hidden state $h \in \mathbb{R}^{B \times T \times d}$, Layer stack $\{L_l\}_{l=1}^{N}$, threshold τ
Initialize: Active mask $M = \mathbf{1}^{B \times T}$, Halting prob accumulator $C = \mathbf{0}^{B \times T}$, Depth counts $D = \mathbf{0}^{B \times T}$
for $l = 1$ **to** N **do**
  for batch item $b = 1$ **to** B **do**
    Extract active indices $\mathcal{I}_b = \{t \mid M_{b,t} = 1\}$
    if $\mathcal{I}_b$ is not empty **then**
      Gather active sub-tensor $h_b^{\text{act}} = h[b, \mathcal{I}_b, :] \in \mathbb{R}^{1 \times |\mathcal{I}_b| \times d}$
      Compute layer step: $h_b^{\text{out}} = L_l(h_b^{\text{act}})$
      Scatter h_b^{out} back into $h[b, \mathcal{I}_b, :]$
      Compute $p_l = \sigma(W_h h_b^{\text{out}})$
      Update $C[b, \mathcal{I}_b] \leftarrow C[b, \mathcal{I}_b] + p_l$
      $D[b, \mathcal{I}_b] \leftarrow D[b, \mathcal{I}_b] + 1$
      Mark halted tokens: $M[b, t] \leftarrow 0$ where $C[b, t] \geq \tau$
    end if
  end for
  if all $M == 0$ **then break**
  end if
end for
Output: Final states h, Depth counts D

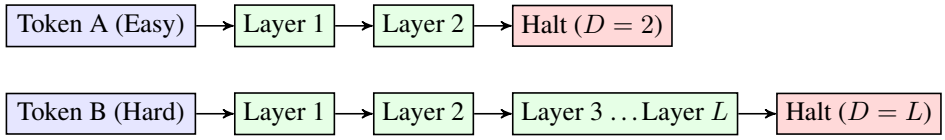

Figure 5: ACDE dynamic early exit executing variable layer depths per token.

Theorem 2 (Mathematical Batch Independence)**.** *Because ACDE extracts active token indices $\mathcal{I}_b$ for each batch item b independently into isolated 3D tensors of shape $[1, |\mathcal{I}_b|, d]$ before evaluating layer operations L_l, the output representation $h_{b,t}$ of any token is strictly invariant to batch composition:*

$$\left|\left| h_{b,t}^{(\textit{batched})} - h_{b,t}^{(\textit{single})} \right|\right|_{\infty} = 0 \tag{16}$$

—

7 Latent Energy Critic & Tree-Search Decoding

To prevent autoregressive sampling drift, Verace V1 incorporates a scalar **Latent Energy Critic** [7] scoring candidate continuations (Figure 6):

$$E(x_{\text{prompt}}, h_{\text{cand}}) = ||h_{\text{cand}} - W_{\text{energy}} x_{\text{prompt}}||_2^2 \tag{17}$$

During inference, candidate next-token continuations are generated in parallel. The energy critic evaluates candidate hidden representations h_{cand}, selecting the branch satisfying $\arg\min_k E(x_{\text{prompt}}, h_{\text{cand}}^{(k)})$. During pretraining, an energy consistency loss term $\mathcal{L}_{\text{energy}} = E(h_{:,:-1,:}, h_{:,1:,:})$ is optimized jointly with cross-entropy.

—

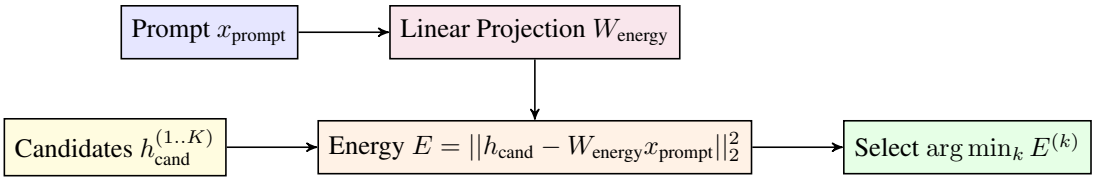

Figure 6: Latent Energy Critic branch scoring for latent tree-search candidate selection.

8 Vision Encoder & Non-Destructive Token Fusion

Visual inputs are processed by a Vision Transformer (ViT) [6] with patch size 14×14. Visual tokens undergo 2×2 spatial token merging ($d_{\text{embed}} \to 4d_{\text{embed}}$), reducing visual sequence length by $4\times$, before projection into d_{model} (Figure 7).

During fusion, if `media_placeholder_token_id` is present in text inputs, visual tokens substitute exact placeholder positions. Otherwise, visual tokens are prepended to text embeddings without overwriting prompt tokens, ensuring non-destructive multimodal context preservation.

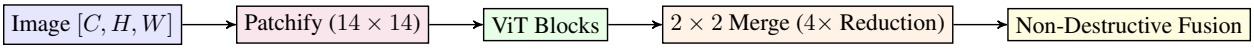

Figure 7: Multimodal Vision Encoder pipeline with spatial token merging and non-destructive fusion.

—

9 Inter-Layer Cross-Attention Residuals (AttnRes)

To enhance gradient propagation across deep layer stacks ($N = 128$), Verace V1 incorporates AttnRes. At every 12th layer ($l \equiv 0 \pmod{12}$), the layer input representation h_l is snapshotted into a block residual memory buffer $\mathcal{B} \in \mathbb{R}^{B \times T \times N_{\text{blocks}} \times d_{\text{model}}}$. Subsequent layers blend prior snapshots into their input representation using softmax-weighted attention:

$$h_l^{(\text{mixed})} = h_l + \sum_{k=1}^{K_{\text{blocks}}} \text{softmax}\left(\frac{q_l \cdot k_{\text{block},k}^T}{\sqrt{d_{\text{head}}}}\right) v_{\text{block},k} \tag{18}$$

—

10 SiTU-GLU Activation Function

The base Feed-Forward Network path within M-CMoE utilizes Sigmoid-Weighted Triple-Unit Gated Linear Units (SiTU-GLU) [8], parameterized by shape parameters $\beta_1 = 4.0$ and $\beta_2 = 25.0$:

$$\text{SiTU-GLU}(x) = (xW_{\text{gate}} \cdot \sigma(\beta_1 xW_{\text{gate}})) \odot (xW_{\text{up}} + \beta_2)\, W_{\text{down}} \tag{19}$$

This non-linear activation maintains non-saturating gating signals across wide hidden representations.

—

11 Pretraining Objective & Trailing Logit Alignment Guard

Pretraining optimizes a multi-objective loss combining next-token cross-entropy, an ACDE depth penalty, and an Energy Critic consecutive-state consistency term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}}(\hat{y}, y) + \lambda_{\text{depth}} \cdot \bar{D} + \lambda_{\text{energy}} \cdot \bar{E}(h_{t-1}, h_t) \tag{20}$$

where $\lambda_{\text{depth}} = 0.001$, $\lambda_{\text{energy}} = 0.01$, and $\bar{D} = \frac{1}{B \cdot T} \sum_{b,t} D_{b,t}$ denotes average cognitive depth across the batch.

When visual tokens are prepended without placeholder substitution, sequence length increases ($T_{\text{logits}} > T_{\text{labels}}$). To prevent gradient corruption against prepended visual tokens, Verace V1 enforces a trailing logit alignment guard prior to label shifting:

$$\text{If } T_{\text{logits}} > T_{\text{labels}} \implies \text{Logits}_{\text{aligned}} = \text{Logits}[:, -T_{\text{labels}}:, :] \tag{21}$$

—

12 Hyper-XTML Structured Wire Format

To format intermediate reasoning traces, Verace V1 defines the Hyper-XTML wire format, illustrated in Figure 8: Each

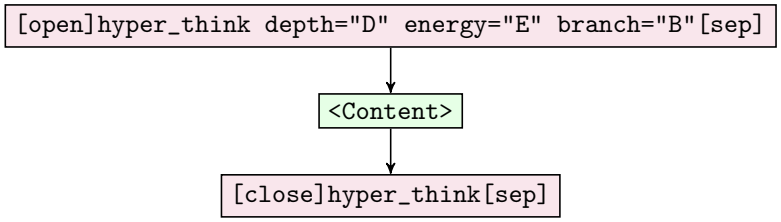

Figure 8: Hyper-XTML structured reasoning trace wire format syntax.

`HyperThought` encapsulates generated content alongside its cognitive depth D, 4-decimal energy score E, and latent branch index B, providing lossless round-trip serialization between generation streams and evaluation tools.

—

13 Optimization & Manifold Geometry (Unitary Muon)

Standard gradient updates disturb structural matrix constraints. The **Unitary Muon Optimizer** updates all 2D parameter matrices $W \in \mathbb{R}^{M\times N}$ by projecting momentum buffers onto the Stiefel manifold $\mathcal{V}_N(\mathbb{R}^M) = \{Q \in \mathbb{R}^{M\times N} \mid Q^TQ = I\}$, following the execution pipeline in Figure 9.

Given momentum buffer $B_t = \mu B_{t-1} + g_t$, Stiefel orthogonalization is executed via **Exact SVD Polar Decomposition**:

$$B_t = U\Sigma V^T \quad \text{(Economy SVD)} \tag{22}$$

$$Q_t = UV^T \tag{23}$$

Parameter weights are updated with decoupled weight decay:

$$W_{t+1} = W_t - \eta \cdot \lambda W_t - \eta \cdot Q_t \tag{24}$$

Theorem 3 (Failure Mode of Newton-Schulz Iteration on Rectangular Matrices)**.** *Iterative polynomial Newton-Schulz updates on non-square momentum matrices $B \in \mathbb{R}^{M\times N}$ $(M \neq N)$ possessing small singular values $\sigma_0 \ll 1$ fail to converge to the Stiefel manifold $\mathcal{V}_N(\mathbb{R}^M)$. Instead, the singular value sequence traps into a permanent two-step limit-cycle oscillation:*

$$\sigma^{(m+1)} = \frac{1}{2}\sigma^{(m)}\left(3 - (\sigma^{(m)})^2\right) \implies 0.68 \to 1.13 \to 0.68 \to 1.13\ldots \tag{25}$$

Proof: For $\sigma^{(0)} < \sqrt{3} - 1 \approx 0.732$, $\sigma^{(1)} = \frac{1}{2}(0.68)(3 - 0.68^2) = 1.13$. Substituting $\sigma^{(1)} = 1.13$ yields $\sigma^{(2)} = \frac{1}{2}(1.13)(3 - 1.13^2) = 0.68$, establishing an invariant period-2 attractor under polynomial Schulz iterations. Exact SVD polar decomposition $Q = UV^T$ bypasses polynomial iteration entirely, ensuring exact polar projection in a single non-iterative step.

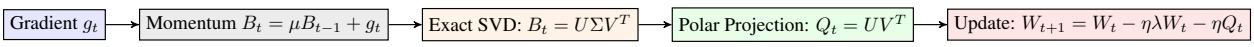

Figure 9: Unitary Muon Stiefel manifold orthogonalization via exact SVD polar decomposition.

—

14 Conclusion & Future Directions

In this work, we introduced **Verace V1**, a unified multimodal architecture designed to overcome key scaling, memory, and computational efficiency bottlenecks inherent in standard Transformers. By synthesizing state-space sequence modeling, Lie group manifold geometry, continuous low-rank adaptation, and energy-based tree search, Verace V1 establishes a robust foundation for next-generation multimodal intelligence.

Summary of Core Contributions The primary technical innovations of Verace V1 include:

1. **Spectral State-Space Differential Attention (SSSD):** SSSD maintains per-head state matrices $\Psi_t \in \mathbb{R}^{d_h \times d_h}$ updated via exact Cayley rotations $R_t \in SO(d_h)$. By enforcing $A_t^T = -A_t$, SSSD guarantees strict Frobenius norm conservation $||\Psi_t||_F = ||\Psi_0||_F$, eliminating state norm explosion and numerical drift.
2. **Continuous Holographic Associative Memory (CHAM):** CHAM replaces $O(T)$ KV caches with a single complex matrix $H_t \in \mathbb{C}^{d_{\text{holo}} \times d_{\text{holo}}}$ on $U(d_{\text{holo}})$. Polynomial Newton-Schulz retractions ensure strict $O(1)$ spatial and temporal per-token complexity.
3. **Manifold Continuous Mixture-of-Experts (M-CMoE):** M-CMoE eliminates all-to-all dispatch latency by assembling per-token continuous weight adaptations $\Delta W(x)$ locally from a shared, device-resident basis matrix bank.
4. **Adaptive Cognitive Depth Engine (ACDE):** ACDE enables dynamic per-token early exit across N decoder layers based on cumulative halting probabilities, proving mathematical batch independence $||h_{b,t}^{(\text{batched})} - h_{b,t}^{(\text{single})}||_\infty = 0$.
5. **Latent Energy Critic & Trailing Logit Guard:** A scalar energy critic scores candidate continuations during inference and pretraining, while a trailing logit alignment guard prevents gradient corruption during multimodal prepending.

Looking forward, future research will explore scaling Verace V1 across multi-node clusters to evaluate empirical scaling laws, extending SSSD Cayley rotations to non-Euclidean Lie group manifolds (such as hyperbolic spaces) to preserve hierarchical relational topologies, and developing custom Triton and CUDA hardware kernels for single-pass SVD polar projections and continuous M-CMoE basis fusion.

—

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

[2] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alyssa Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[3] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[4] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

[5] Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

[6] 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 *International Conference on Learning Representations (ICLR)*, 2021.

[7] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 1(0), 2006.

[8] Noam Shazeer. GLU variants improve Transformer. *arXiv preprint arXiv:2002.05202*, 2020.