Semantic Routing: Exploring Multi-Layer LLM Feature Weighting for Diffusion Transformers

Bozhou Li[1,2,*] Yushuo Guan[2,†] Haolin Li[3] Bohan Zeng[1] Yiyan Ji[4] Yue Ding[5] Pengfei Wan[2] Kun Gai[2] Yuanxing Zhang[2,‡] Wentao Zhang[1]

[1]Peking University [2]KlingAI [3]Fudan University [4]Nanjing University [5]UCAS | *Intern at KlingAI (libozhou@pku.edu.cn) †Project Leader ‡Corresponding Author

Peking University · KlingAI

NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2026 · Main Track Poster

01 Motivation: Static Interface vs. Hierarchy

Modern DiTs pair with expressive decoder-only LLMs, yet still reuse **one fixed text layer** across every denoising step t and block d.

LLM Layer Supply ($l \in [1, L]$)

- **Shallow ($l \ll L$):** lexical tokens, local color & surface attributes.
- **Middle:** spatial/counting bindings & attenuated attention sinks.
- **Deep ($l \approx L-1$):** conceptual abstractions & global layout priors.

DiT Evolving Demand (t, d)

- **Flow Time $t \in [0,1]$:** coarse global layout early → fine texture late.
- **Block Depth $d \in [0, D-1]$:** structural formation in shallow blocks vs. detail synthesis in deep blocks.

Central Research Question ⟶ When multi-layer LLM features are available, should conditioning route across **flow time** t, **DiT depth** d, or **both** under strictly matched training?

02 Normalized Convex Routing Interface

Given frozen hidden states $\{H^{(l)}\}_{l=1}^{L}$, **Semantic Routing** builds the text condition at site (t, d) on the simplex:

$$H_{\text{cond}}(t,d) = \sum_{l=1}^{L} \alpha_{t,d}^{(l)} \, \mathrm{LN}(H^{(l)}), \;\; \alpha_{t,d} = \mathrm{Softmax}(z_{t,d}) \in \Delta^{L-1}$$

Pre-Fusion LayerNorm
Eliminates severe cross-layer norm growth so high-norm layers cannot hijack Softmax routing.

Convex Hull Stability
$\sum_l \alpha_{t,d}^{(l)} = 1$ bounds feature variance across all DiT cross-attention blocks.

Zero-Init Time Gate (TCFG)
Sinusoidal $\phi(t)$ + 2-layer SiLU MLP, zero-init to start at uniform $\mathbf{1}/L$.

Flexible DiT K/V Projection
Preserves DiT freedom to re-project routed tokens into K/V (vs. FuseDiT bottleneck).

03 Design Space: 6 Controlled Routing Schemes

Matched Protocol: 2.24B DiT (D=24) · Qwen3-VL-4B (L=36) · 30M Captions · 500k Steps

Variant	Routing Logits	t	d	Empirical Behavior
B1: Penult.	$\mathrm{LN}(H^{(L-1)})$	–	–	Underutilizes LLM hierarchy
B2: Uniform	$\frac{1}{L}\sum_l \mathrm{LN}(H^{(l)})$	–	–	Strong stationary prior
B3: Static	$z_{t,d} = \beta \in \mathbb{R}^L$	–	–	Optimization drift (< B2)
S1: Time	$z_{t,d} = g_\psi(\phi(t))$	✓	–	Trajectory lag → blur
S2: Depth	$\mathbf{z_{t,d}} = \boldsymbol{\beta_d} \in \mathbb{R}^{\mathbf{L}}$	–	✓	Best alignment & fidelity
S3: Joint	$z_{t,d} = g_{\psi_d}(\phi(t))$	✓	✓	Depth dampens temporal drift

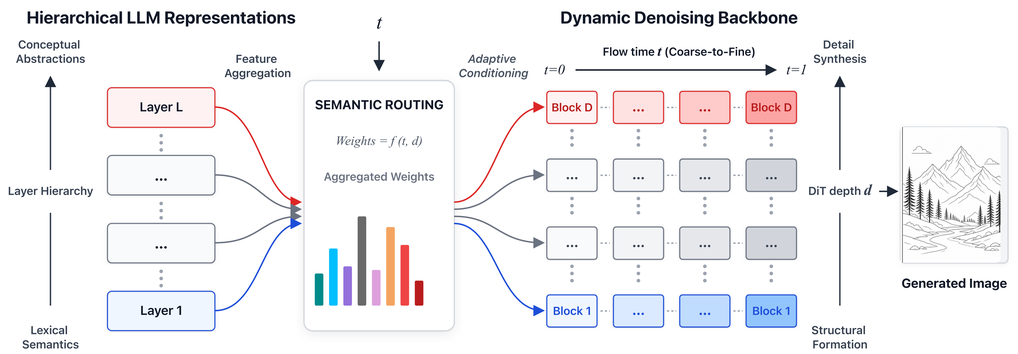
Supply-Demand Relaxation ($S_{t,d} = Q\alpha_{t,d}$)

Global β (**B3**) chases shifting denoising stages into a suboptimal compromise, whereas per-block β_d (**S2**) relaxes global sharing across D=24 blocks (864 scalars):

$$\min_{\alpha \in \Delta^{L-1}} \sum_{d=0}^{D-1} \|R_d - Q\alpha\|_2^2 \;\geq\; \sum_{d=0}^{D-1} \min_{\alpha_d \in \Delta^{L-1}} \|R_d - Q\alpha_d\|_2^2$$

▸ **Takeaway:** DiT depth d is the minimal, reliable axis for multi-layer LLM routing.

04 Architecture Overview: Semantic Routing in Diffusion Transformers



1. Frozen LLM Hierarchy
Shallow lexical/color → mid spatial/counting → deep global layout across L=36 layers.

2. Simplex Semantic Router
LayerNorm + Softmax weights $\alpha_{t,d} \in \Delta^{L-1}$ synthesize norm-bounded conditioning.

3. Per-Block DiT Injection
Each DiT cross-attention block $d \in [0, D-1]$ queries its stage-matched semantic mixture.

05 Quantitative Benchmarks: Depth-Wise Routing Leads Alignment

+9.97 GenAI Counting — 74.57 (S2) vs. 64.60 (B1)

+8.92 Differentiation — 74.31 (S2) vs. 65.39 (B1)

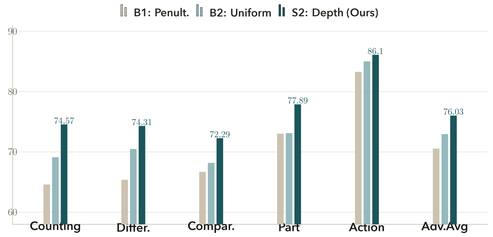
79.07 GenAI-Bench Avg — Rank #1 (+4.11 over B1)

67.07 GenEval Overall — Rank #1 (Reward 3.06)

Main Results (500k Steps, Table 1)

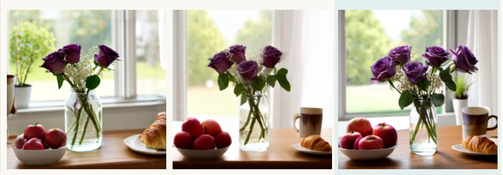
Method	GenEval ↑	GenAI ↑	Reward ↑	FID ↓
B1: Penult.	64.54	74.96	3.02	25.52
B2: Uniform	66.51	76.82	**3.06**	**24.77**
B3: Static	64.77	76.31	3.05	25.25
FuseDiT (K/V)	60.95	75.02	3.05	24.86
S1: Time-wise	63.41	76.20	2.97	24.95
S2: Depth	**67.07**	**79.07**	**3.06**	24.88
S3: Joint	66.05	77.44	**3.06**	25.24

GenAI-Bench Fine-Grained Skills



06 Qualitative Comparison: Compositional Counting & Spatial Grounding

Case 1 (Counting): "A vase with five purple roses on a kitchen table."



× 3 roses | × 4 roses | ✓ Exact 5 roses

Case 2 (Spatial): "Three ceramic cups sit to the right of a wooden fork."

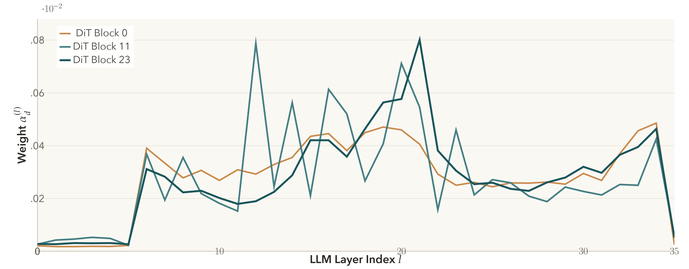


× Missing fork | × 4 cups | ✓ Fork left + 3 cups

Why S2 Excels Visually: Middle LLM layers supply explicit object bindings to early DiT blocks, while shallow lexical layers refine late surface details.

07 Learned Depth-Wise Routing Weights

Softmax weights $\alpha_d^{(l)}$ across LLM layers (S2)



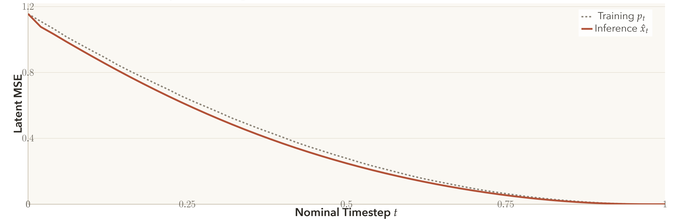
Layer Specificity
Boundary layers (0,35) get ≈ 0 weight; L−1 anchors early layout then fades.

Neighbor Inhibition
Softmax suppresses correlated residual neighbors, selecting sparse local representatives.

Emergent Locality
Adjacent DiT blocks show smooth JS similarity without explicit cross-block regularization.

08 Why Time-Only Fails: Trajectory Lag

Latent MSE: inference $\hat{x}_t$ outpaces training schedule p_t



Semantic Lag ($\tau(t) > t$): Recursive ODE states $\hat{x}_t$ outpace static p_t, so $g_\psi(t)$ injects coarse priors into clean latents. Advancing $t' = \mathrm{clip}(t + \delta(t))$ partially recovers S1:

Table 3	GenEval	GenAI	Rew.
S1	63.41	76.20	2.97
S1+Shift	**63.65**	**76.46**	**2.98**
S3 / +Shift	66.05/66.00	77.44/**77.65**	3.06

09 Cross-Encoder Scaling & Practical Guidelines

T5Gemma-2B + 3.95B DiT

Encoder-decoder LLM (8T tokens, 250k steps):

Method	GenEval	GenAI	Style	FID ↓
Baseline	36.58	42.74	2.89	31.76
Norm-Avg	34.68	41.64	2.83	30.23
Depth (S2)	41.12	50.05	**2.89**	29.40

+4.54 GenEval & +7.31 GenAI even where Norm-Avg fails.

Inference Latency (H800, `compile`)

Res.	B1: Penult.	S2: Depth	Δ
256×256	491.6 ms	547.7 ms	+56 ms
512×512	1176.2 ms	1241.3 ms	**+5.5%**

01 Default to Depth (β_d)
Unlocks multi-layer LLM semantics reliably across decoder-only & encoder-decoder LLMs with 864 scalars.

02 Pre-Normalize (LN)
Eliminates cross-layer hidden-state norm spikes before Softmax mixing to stabilize co-adaptation.

03 Calibrate Time (t)
Nominal t lags behind recursive ODE denoising; trajectory-aware SNR alignment is essential before time gating.

KEY TAKEAWAYS ■ **DiT Depth d is the Reliable Axis:** Per-block routing (**S2**) gains **+9.97** Counting & **+4.11** GenAI-Bench over B1. ■ **Temporal Lag:** Iterative ODE outpaces nominal t ($\tau(t) > t$), blurring time-only gates. ■ **Generalizable:** Holds across **Qwen3-VL** & **T5Gemma**.