LAYERBRIDGE: UNIFYING UNDERSTANDING, GENERATION AND EDITING WITH ADAPTIVE ATTENTION OVER FROZEN VLM LAYERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent approaches to unified multimodal modeling use pretrained vision-language models (VLMs) to encode prompts or noisy visual representations as conditions for a generation decoder. However, these methods typically condition the decoder on final-layer features or on fixed combinations of multi-layer prompt features, limiting each decoder block's ability to select the layers it needs as denoising progresses. Motivated by this, we propose LayerBridge, which uses depth attention to let a pixel-diffusion decoder dynamically select and combine features from multiple layers of a frozen VLM's vision encoder and language model, reusing its learned representations while obtaining multi-granularity guidance. The decoder uses its current state and the denoising timestep to weight and combine these features, allowing different mixtures across blocks, image regions and noise levels. This adaptive readout also provides a general architecture for feature reuse across generation, understanding, and editing. For understanding, the same mechanism lets visual tokens within the VLM's language model query features from multiple vision-encoder layers, providing finer visual information for vision-centric perception. For editing, reference tokens retrieve fine visual cues from intermediate vision layers and pass them to target states through LLM attention, preserving reference details during editing. To learn this shared interface, we keep the VLM frozen while training the added modules across generation, understanding, and editing. With Qwen3.5-0.8B/4B and 0.3B/1.1B generation-specific parameters, LayerBridge reaches **0.87/0.90** on GenEval, **0.54/0.62** on WISE, and **3.86/4.16** on ImgEdit, together with gains in vision-centric perception over native Qwen3.5.

1 INTRODUCTION

Pretrained vision-language models (VLMs) offer visual knowledge and reasoning that can guide image synthesis. Connecting them to diffusion decoders provides a practical route to unified multimodal modeling. MetaQuery (Pan et al., 2025) uses learnable queries to transfer conditions from a frozen VLM while preserving its understanding capabilities. To make this guidance responsive to the evolving image, RepFusion (Pan et al., 2026) and UniDDT (Wang et al., 2026c) also encode the noisy target with the VLM at each denoising step. Their decoders process noisy latents directly, while VLM features relate the current target to the prompt. Re-encoding updates this guidance over time, but these decoders can read only final-layer LLM features.

The choice of final outputs is natural for a VLM trained to predict text, but image synthesis can benefit from information represented earlier in the network. Prior VLM studies find that shallow layers retain visual details that complement the contextual representations of deeper layers (Li et al., 2026d). Our experiments on frozen Qwen3.5 (Qwen Team, 2026) show that reconstruction and local grounding favor different layers (Figure 2). These complementary readouts motivate giving the decoder direct access to multiple VLM layers beyond the final output.

However, how to exploit these layers remains an open question. UniFusion (Li et al., 2025) and Semantic Routing (Li et al., 2026a) explore multi-layer conditioning, but they combine only the layers of the input prompt rather than those of the image being generated, and their layer mixtures do not adapt to the current decoder state. MoDA (Zhu et al., 2026a) and Attention Residuals (Kimi

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

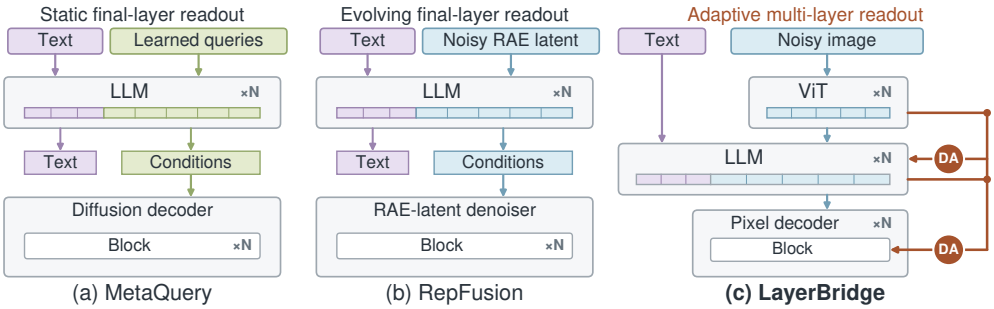


Figure 1: **VLM readout paths.** MetaQuery uses fixed final-layer LLM conditions. RepFusion re-encodes the noisy target to update these conditions each step. In LayerBridge, depth attention (DA) lets LLM image tokens read intermediate visual-encoder layers and decoder states read target-image features across vision and LLM layers. Depth selection uses the current image-token state and timestep. For clarity, we omit the baselines' noisy-latent inputs and the LayerBridge connector.

Team et al., 2026) address layer reuse within one network, letting later blocks attend to earlier representations instead of inheriting a fixed accumulation of outputs, so that each block can retrieve the representation its own computation needs. Inspired by this principle, we make depth selection a cross-network interface, where each decoder block uses its current image-token state and the diffusion timestep to select features of the re-encoded target at the same image region.

We introduce LayerBridge, a unified multimodal model that applies this cross-network depth attention to generation, understanding and editing (Figure 1). Selected decoder blocks combine intermediate vision-encoder and LLM image features according to their current states and timesteps. This lets each block draw on visual detail and prompt-conditioned context at corresponding image positions. Because the pretrained vision encoder takes RGB images, we use pixel diffusion to keep the evolving target in its input space and re-encode the target with the VLM at every denoising step.

Access to visual detail also matters when recognizing object attributes or preserving them during an edit. We apply the same query principle between ViT and LLM, letting LLM image tokens retrieve intermediate vision features for understanding. For editing, reference tokens read their own visual features, and LLM attention transfers these cues to target states under the instruction. The decoder reads the resulting target features, preserving a shared generation path while giving reference details a route into the edited image. These reference reads share their vision-to-LLM weights with understanding reads, linking detail preservation and perception during joint training.

Training first freezes both the VLM and the MiniT2I-initialized decoder (Wang et al., 2026d), learning depth attention and the connector. It then adds editing and unfreezes the decoder for joint training, retaining understanding supervision and the frozen VLM. With Qwen3.5-0.8B/4B backbones and 0.3B/1.1B generation-specific parameters, each scale supports generation, understanding and editing in one model. The smaller model keeps GenEval at the level of its MiniT2I decoder while raising WISE from 0.42 to **0.54** over MiniT2I trained on our data. The larger model reaches **0.90** on GenEval, **0.62** on WISE and **4.16** on ImgEdit. Shared vision-to-LLM reads improve vision-centric perception over native Qwen3.5, while general benchmarks remain close to native scores.

Our main contributions are as follows.

- **Depth attention for unified multimodal modeling.** We propose LayerBridge, which unifies understanding, generation and editing by selecting task-relevant features from a frozen VLM hierarchy. Shared cross-network depth attention supplies the visual granularity each task needs.
- **Adaptive layer selection across tasks and blocks.** We introduce state- and time-conditioned depth queries that let individual blocks select features of the evolving target for their current computation, rather than fixing one layer mixture for the prompt. Layer selection responds to task demands and local image content, while time conditioning accounts for the changing noise level of the target during synthesis.
- **Competitive performance at compact scale.** Across both backbone scales, LayerBridge combines competitive generation and editing with gains in vision-centric perception over native Qwen3.5, using 0.3B/1.1B generation-specific parameters.

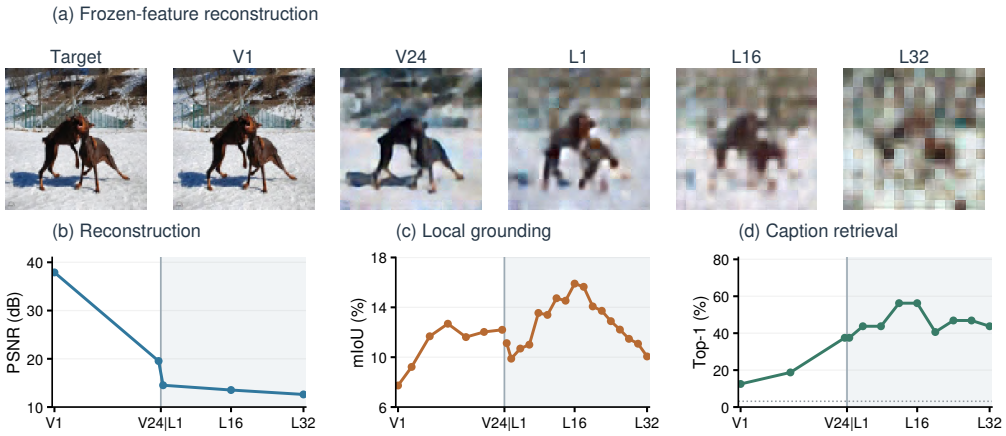


Figure 2: **Complementary information across the pretrained hierarchy.** The top row reconstructs the same photo from Vn/Ln (ViT/LLM layer n). The curves measure reconstruction, local grounding and caption-conditioned retrieval. Shallow vision features retain appearance detail. Intermediate LLM layers improve grounding over the final layer. Intermediate LLM states also depend strongly on the supplied caption. Protocols appear in Section E.

2 Related Work

Unified multimodal understanding and generation. Unified models combine language and image objectives (Chameleon Team, 2024; Wang et al., 2026e; Xie et al., 2025a; Zhou et al., 2025; Xie et al., 2025b; Yang et al., 2025a) through shared representations (Wu et al., 2025c; Qu et al., 2025; Yan et al., 2026; Diao et al., 2026) or task-specific modules (Wu et al., 2025b; Li et al., 2026e; Liu et al., 2026; Shen et al., 2026; Li et al., 2026c). However, one representation may not meet the granularity needs of every task. Mixture-of-transformer models pair each generation layer with the understanding layer at the same depth (Shi et al., 2025; Deng et al., 2025), adding generation weights across the full backbone. LayerBridge keeps the VLM frozen and lets each decoder block retrieve task-relevant features from VLM layers that offer complementary detail and context (Li et al., 2026d; Singh et al., 2026).

Connecting VLMs to image decoders. Learnable queries connect VLMs to image decoders (Dong et al., 2024; Pan et al., 2025; Chen et al., 2025a), BLIP3o-NEXT conditions diffusion on autoregressive hidden states (Chen et al., 2025b), Bifrost-1 predicts patch-level CLIP latents (Lin et al., 2025b), and editing models also encode references to guide changes (Tang et al., 2026; Liu et al., 2025; Wu et al., 2025a). RepFusion and UniDDT re-encode noisy targets to update guidance during denoising (Pan et al., 2026; Wang et al., 2026c). However, their final-layer guidance does not expose intermediate representations. LayerBridge exposes these intermediate features to guide spatial detail and to preserve reference appearance.

Selecting intermediate features. Intermediate features support perception and reconstruction (Huang et al., 2017; Lin et al., 2017; Yao et al., 2024; Meng et al., 2024; Tong et al., 2024a; Singh et al., 2026; Wei et al., 2025; Zhu et al., 2026b). Within one network, dense and hyper-connections mix earlier layers (Pagliardini et al., 2024; Zhu et al., 2025; Xiao et al., 2025), Attention Residuals, MoDA and Depth-Attention attend to earlier states (Kimi Team et al., 2026; Zhu et al., 2026a; Zeng et al., 2026), and LaTtE-Flow and DAR adapt layers within denoisers (Shen et al., 2025b; Xu et al., 2026). For VLM conditioning, UniFusion pools layers (Li et al., 2025) and Semantic Routing weights LLM text layers (Li et al., 2026a), while MammothModa2 and DeepGen average or concatenate layers (Shen et al., 2025a; Wang et al., 2026a). TBAC-UniImage, Playground v3 and HiDream-I1 fix layer-to-block assignments (Xu et al., 2025; Liu et al., 2024a; Cai et al., 2025), and ELLA adapts final LLM features to the timestep (Hu et al., 2024). However, these methods mix conditioning-input layers. LayerBridge instead queries the re-encoded target's layers with the decoder state and timestep, matching depth mixtures to image regions (Table 13).

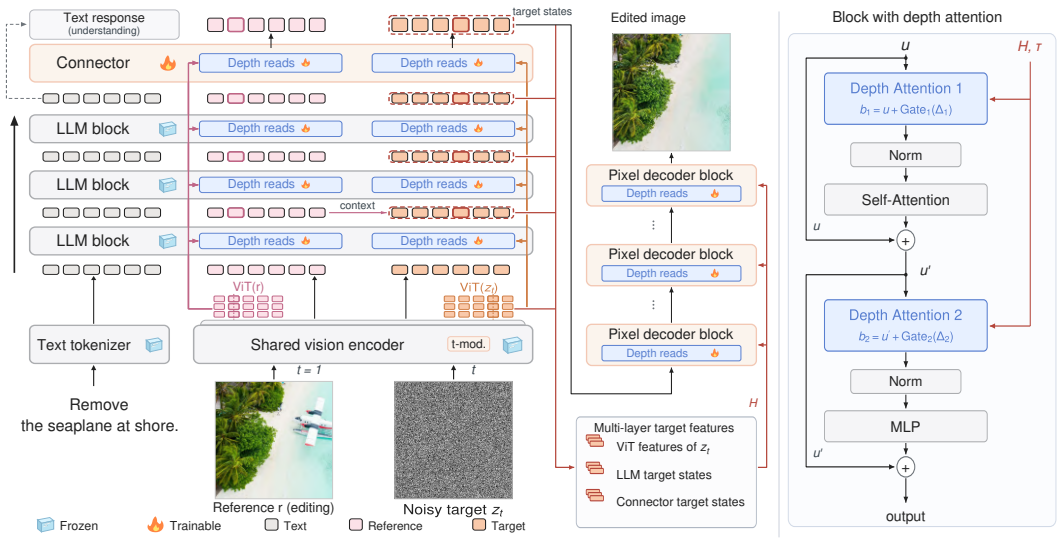


Figure 3: **LayerBridge architecture with an editing example.** The VLM processes an instruction, a clean reference and a noisy target. Depth attention lets image-token states read intermediate vision features. The pixel decoder reads target features from vision, LLM and connector layers. Sequence attention incorporates reference information into target states. Generation omits the reference, while understanding uses the VLM text head. In the inset, $\mathbf{u}$ and $\mathbf{u}'$ are states before and after attention, $\mathbf{H}$ collects position-wise feature banks, and τ is the timestep for that image. $\text{Gate}(\Delta) = g(\Delta) \odot \Delta$ as in Equation (2). Section A specifies source layers and decoder read sites.

Pixel diffusion. Latent diffusion compresses images with a VAE or RAE (Rombach et al., 2022; Zheng et al., 2026). Pixel-space models instead use multi-scale flows, simplified scaling or frequency-decoupled decoders (Chen et al., 2025d; Hoogeboom et al., 2025; Wang et al., 2026b; Ma et al., 2026b; Yu et al., 2026), while JiT enables large-patch synthesis through clean-image prediction (Li & He, 2026) and MiniT2I adds text conditioning (Wang et al., 2026d). Pixel targets can enter the VLM encoder directly, so LayerBridge reuses it for noisy targets without a separate latent encoder, adapting it to noise while keeping its clean-image mapping.

3 Method

3.1 Model Overview

LayerBridge pairs a frozen VLM with a MiniT2I-initialized pixel decoder (Wang et al., 2026d). The LLM interprets image content, while the decoder combines this guidance with noisy-target features through depth attention (Figure 3). A trainable connector maps prompt-conditioned LLM image tokens into the decoder's input space. The decoder sees the noisy target only through VLM features, whose shallow vision layers retain finer appearance detail (Figure 2), and generation and editing share all of these connections.

We write a noisy target as $\mathbf{z}_t = t\mathbf{x} + (1-t)\boldsymbol{\epsilon}$, with clean pixels $\mathbf{x}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The timestep $t \in [0, 1]$ runs from pure noise to the clean image (Lipman et al., 2023; Liu et al., 2023).

3.2 Cross-Network Depth Attention

Unlike final-layer, pooling and fixed-assignment interfaces, LayerBridge lets decoder blocks retrieve features for their current states and timesteps as denoising progresses.

In Figure 3, ViT, LLM and connector outputs form a bank indexed by image position and layer. VLM weights remain frozen, but VLM states incorporate modulation and upstream depth reads. LLM and connector image-token states query ViT features, whereas decoder states query target features from all three networks. Qwen3.5 (Qwen Team, 2026) merges 2×2 vision patches into each LLM image

token. Aligning these grids restricts depth attention to corresponding image regions (Section A). Each read updates its querying state without modifying the bank.

Depth Attention. Access to the same bank does not imply that every block or image region needs the same features. We therefore form queries from current image-token states, allowing LLM, connector and decoder blocks to weight layers according to their content and task context. Timestep conditioning additionally identifies the noise level during synthesis, allowing guidance to change from structure formation to detail refinement. Understanding images and editing references use the clean timestep while retaining adaptation through their image-token states.

We adapt depth attention (Zhu et al., 2026a; Kimi Team et al., 2026) from within-network reuse to a cross-network interface. Instead of reusing previously projected keys and values as in MoDA, each read uses its own projections of the aligned bank. At image position p, let $\mathbf{u}_p \in \mathbb{R}^d$ be the block state and $\mathbf{H}_p \in \mathbb{R}^{K\times d}$ the aligned bank. Targets use $\tau = t$ and clean images $\tau = 1$. LLM and connector reads use $\mathbf{e}_\tau = \psi(\tau)$, while decoder reads use a separate embedding. The affine maps A project queries, keys and values. With state and bank normalization R and head normalizations R_q, R_k, width-d_h heads select over K layers using $\bar{\mathbf{H}}_p = R(\mathbf{H}_p)$.

$$\begin{aligned}
\bar{\mathbf{u}}_p &= R(\mathbf{u}_p) + A_\tau(\mathbf{e}_\tau), \quad \mathbf{q}_h = R_q(A_{q,h}(\bar{\mathbf{u}}_p)), \\
\mathbf{K}_h &= R_k\big(A_{k,h}(\bar{\mathbf{H}}_p)\big), \quad \mathbf{V}_h = A_{v,h}(\bar{\mathbf{H}}_p), \\
\Delta_p &= A_o\Big(\mathrm{concat}_h\Big[\mathrm{softmax}_K\Big(\mathbf{q}_h\mathbf{K}_h^\top/\sqrt{d_h}\Big)\mathbf{V}_h\Big]\Big).
\end{aligned} \tag{1}$$

Different positions can select different depth mixtures, different blocks can interpret the same bank differently, and selection can change with the noise level. Table 3 compares the two query inputs.

Gated Integration. To retain pretrained capabilities while learning this interface, we separate depth reads from sequence attention, unlike MoDA's joint attention over sequence and depth. A zero-initialized gate starts adaptation from the pretrained block computation.

$$\begin{aligned}
g(\Delta) &= \mathbf{W}_g\Delta + \mathbf{b}_g, \\
\mathbf{b}_p &= \mathbf{u}_p + g(\Delta_p)\odot\Delta_p, \qquad \mathbf{u}' = \mathbf{u} + F(\mathbf{b}).
\end{aligned} \tag{2}$$

Here $\mathbf{b}_p = \mathbf{u}_p$ outside image positions, and F includes the original sublayer normalization. With $\mathbf{W}_g = \mathbf{b}_g = 0$, the block initially computes $\mathbf{u} + F(\mathbf{u})$. Training learns the read strength. Separate reads before attention and the MLP let the latter query the updated state (Figure 3, inset). Only image tokens query vision layers, with sequence attention and Gated-DeltaNet (Yang et al., 2025b) carrying retrieved information to text tokens. Algorithm 1 and Section A give details.

3.3 Adapting Visual Features across Noise Levels

The VLM was pretrained on clean images, but the depth interface also needs features of noisy targets. We therefore adapt the vision encoder to noise while preserving its clean-image mapping. We also control the LLM input to favor content interpretation, while decoder depth reads retain noisy-target features for local recovery.

Adapting the vision encoder to noise. Timestep-conditioned AdaLN adapts the encoder to the current noise level. To preserve its pretrained clean-image mapping, we anchor this modulation to zero at the clean endpoint. At each vision-block LayerNorm site ℓ, hidden state $\mathbf{h}$ receives scale $\boldsymbol{\gamma}_\ell(t)$ and shift $\boldsymbol{\beta}_\ell(t)$ given by

$$\begin{aligned}
\mathrm{LN}(\mathbf{h}) &\mapsto \mathrm{LN}(\mathbf{h})\odot\big(1+\boldsymbol{\gamma}_\ell(t)\big) + \boldsymbol{\beta}_\ell(t), \\
[\boldsymbol{\gamma}_\ell, \boldsymbol{\beta}_\ell] &= \mathbf{W}_\ell\,\mathrm{MLP}\big(\phi(t) - \phi(1)\big),
\end{aligned} \tag{3}$$

where ϕ is a sinusoidal timestep embedding. Bias-free MLPs give $\mathrm{MLP}(\mathbf{0}) = \mathbf{0}$, making modulation exactly zero at $t = 1$ throughout training. Clean images therefore retain the pretrained encoder representation expected by the LLM, while noisy targets receive learned corrections. Zero-initialized $\mathbf{W}_\ell$ also preserves every vision block at initialization. Without the anchor, understanding drops by 16.9 on MME perception (MME-P) at unchanged GenEval (Table 15).

Controlling noisy inputs to the LLM. Vision adaptation alone does not determine what should enter the frozen LLM. The LLM is pretrained to interpret recognizable visual content, whereas pixel recovery also requires features of the current noise realization. At low signal-to-noise ratios, routing both through the LLM would burden semantic interpretation with noise-dependent detail. We therefore attenuate its direct visual input, encouraging it to relate the remaining content evidence to the prompt. The pixel decoder still reads position-aligned vision features directly through depth attention, retaining a separate path to noisy-target information for local recovery.

Following the signal coefficient in $\mathbf{z}_t$, we interpolate each merged encoder token $\mathbf{s}_t$ (Qwen Team, 2026) with a learned timestep embedding $\psi(t)$,

$$\tilde{\mathbf{s}}_t = t\mathbf{s}_t + (1-t)\psi(t).$$

This suppresses noise-only tokens at the LLM entrance while retaining the original clean-image input for references and understanding. Clean-token supervision encourages content-bearing inputs (Section 3.5). Table 15 separates attenuation, timestep conditioning and alignment.

3.4 Unified Multimodal Inference

Generation and editing share a denoising process that refreshes VLM guidance as the target develops. At each step, the VLM encodes the noisy target, and the decoder reads the updated feature banks to predict clean pixels $\hat{\mathbf{x}}$ for the sampler. Editing adds a clean reference between the instruction and target. Reference image tokens retrieve vision-layer detail, which sequence attention incorporates into target states. The decoder can thus access reference appearance even at high noise, without pointwise alignment between the images or additional editing parameters.

Understanding reuses the vision-to-LLM depth attention that supplies reference detail for editing. With clean images at $t = 1$, image-token states select vision features according to their context, and the language head produces the response. The native comparison disables depth reads. Sections A and B give the attention masks and sampling settings.

3.5 Training Objectives and Stages

Training supervises both the content entering the frozen LLM and the pixels reconstructed by the decoder. Token attenuation controls input strength but does not determine its content. We therefore align $\tilde{\mathbf{s}}_t$ with stop-gradient features $\mathbf{s}^\star$ from the frozen encoder of the clean target, alongside weighted clean-image prediction (Li & He, 2026).

$$\begin{aligned}\mathcal{L}_{\text{image}} &= \mathbb{E}_{\mathbf{x},t,\boldsymbol{\epsilon}}\left[\frac{w(t)}{N_x}\,\|\hat{\mathbf{x}}-\mathbf{x}\|_2^2\right] + \lambda\mathcal{L}_{\text{align}},\\ \mathcal{L}_{\text{align}} &= \mathbb{E}_{\mathbf{x},t,\boldsymbol{\epsilon}}\left[\frac{\|\tilde{\mathbf{s}}_t-\mathbf{s}^\star\|_2^2}{\max\left(\|\mathbf{s}^\star\|_2^2, D\epsilon_a\right)}\right].\end{aligned} \tag{4}$$

Here N_x and D count pixel and feature entries per image. We use $\lambda = 1$ and $\epsilon_a = 10^{-6}$. Clean-feature energy normalizes the alignment term across images, and Section A defines $w(t)$. Understanding uses next-token cross-entropy, supervising shared depth attention and ψ.

Training first establishes compatible connections between the pretrained VLM and decoder, then adapts the decoder for joint generation, understanding and editing. Stage 1 freezes both pretrained networks and trains depth attention, the connector and image projections on generation and understanding. Stage 2 keeps the VLM frozen, unfreezes the decoder and adds editing. Stage 3 refines all three tasks at a lower learning rate. Understanding supervision constrains the shared interface in every stage, with task schedules and loss scales specified in Section A.

4 Experiments

4.1 Implementation details

Architectures. LayerBridge-0.8B and LayerBridge-4B build on the frozen Qwen3.5-0.8B and Qwen3.5-4B (Qwen Team, 2026), a 17-block pixel decoder initialized from the pretrained MiniT2I

Model	Size	Generation			Understanding		
		GenEval	DPG	WISE	MME-P	MMB	MMMU
Specialized text-to-image models							
SDXL (Podell et al., 2024)	2.6B	0.55	74.7	–	–	–	–
SD3-Medium (Esser et al., 2024)	2B	0.62	84.1	–	–	–	–
FLUX.1-dev (Black Forest Labs, 2024)	12B	0.66	83.8	–	–	–	–
MiniT2I-B/16 (Wang et al., 2026d)	0.3B+0.3B	0.87	84.2	0.40	–	–	–
MiniT2I-L/16 (Wang et al., 2026d)	0.3B+0.9B	0.88	85.9	0.45	–	–	–
Unified multimodal models							
Janus-Pro-1B (Chen et al., 2025e)	1B	0.73	82.6	0.27	–	–	–
Janus-Pro-7B (Chen et al., 2025e)	7B	0.80	84.2	0.33	1567	79.2	41.0
OmniGen2 (Wu et al., 2026)	3B+4B	0.80	83.6	–	–	–	–
BAGEL (Deng et al., 2025)	7B+7B	0.82	85.1	0.35	1687	85.0	55.3
BLIP3-o-8B (Chen et al., 2025a)	7B+1.4B	0.84	81.6	0.52	–	–	–
UniDDT (Wang et al., 2026c)	4B+1B	0.87	86.9	0.58	**1699.5**	82.2	52.6
MammothModa2 (Shen et al., 2025a)	8B+3B+2B	0.87	87.2	–	–	–	–
DeepGen 1.0 (Wang et al., 2026a)	3B+2B	0.87	87.9	0.57	–	–	–
Ovis-U1 (Wang et al., 2025a)	2.4B+1.2B	0.89	83.7	0.39	–	–	–
UniLIP-1B (Tang et al., 2026)	1B+0.6B	0.88	–	0.48	1499	72.6	43.3
UniLIP-3B (Tang et al., 2026)	2B+1.6B	**0.90**	–	0.54	1636	80.7	48.7
Frozen-VLM conditioning							
MetaQuery-L (Pan et al., 2025)	3B+1.6B	0.78†	81.1	–	1574.3	78.6	53.1
MetaQuery-XL (Pan et al., 2025)	7B+1.6B	0.80†	82.1	–	1685.2	83.5	58.6
RepFusion-SFT (Pan et al., 2026)	7B+3.2B+1.6B	0.87	85.1	–	–	–	–
Qwen3.5-0.8B (native)	0.8B	–	–	–	1306.8	68.0	47.4
LayerBridge-0.8B	0.8B+0.3B	0.87	84.7	0.54	1327.5	68.5	47.3
Qwen3.5-4B (native)	4B	–	–	–	1569.1	89.4	**77.6**
LayerBridge-4B	4B+1.1B	**0.90**	**88.2**	**0.62**	1586.3	**89.6**	77.4

Table 1: **Generation and understanding.** Bold marks the best results. Sizes list backbone and generation-specific parameters, and LayerBridge adds 0.1B/0.7B of shared reads. SDXL, SD3 and FLUX.1 list denoiser sizes. † denotes rewritten GenEval prompts. WISE uses the Verified release. Evaluation protocols and baseline sources appear in Sections B and D.

generator (Wang et al., 2026d), and a 6-block connector at LLM width. Generation-specific parameters total 0.3B and 1.1B, and all added modules, including shared reads, total 0.4B and 1.8B (Section A). Both scales read at 6 decoder blocks (read sites), spaced $s_{\mathrm{dec}} = 3$ apart.

Training. Both scales follow the three stages in Section 3.5, using approximately 50k, 100k and 20k updates. Stage 1 aligns the interfaces on 3M pairs from BLIP3o-Pretrain-Short-Caption (Chen et al., 2025a). Stage 2 jointly trains on 3M Fine-T2I generation pairs (Ma et al., 2026a) and 1.1M edits from GPT-Image-Edit-1.5M (Wang et al., 2025c). Stage 3 refines the model on 100k generation pairs from BLIP3o-60k (Chen et al., 2025a) and ShareGPT-4o-Image, alongside 40k edits from the latter (Chen et al., 2025c). Understanding uses 1M LLaVA-OneVision-1.5 examples (An et al., 2025), retaining 50k for Stage 3.

Evaluation. We evaluate generation on GenEval (Ghosh et al., 2023), DPG-Bench (Hu et al., 2024) and WISE (Niu et al., 2026) with raw benchmark prompts, while published baselines retain their source settings. WISE uses the Verified protocol and one judge for all rows. We evaluate editing on ImgEdit (Ye et al., 2025) and GEdit-Bench-EN (Liu et al., 2025). MME (Fu et al., 2025), MMBench-EN (Liu et al., 2024b) and MMMU (Yue et al., 2024) assess understanding through VLMEvalKit (Duan et al., 2024). Each scale uses one joint checkpoint across tasks.

4.2 Comparison with state of the art

Image generation. GenEval reaches 0.87/0.90 at 0.8B/4B, matching RepFusion-SFT and UniLIP-3B, and the 4B model reaches 88.2 on DPG (Table 1 and Figure 4). Trained on our data without a VLM, the MiniT2I-B decoder that initializes both scales reaches 0.86 GenEval but only 0.42 WISE

Model	Size	ImgEdit Overall	GEdit-Bench-EN SC	GEdit-Bench-EN PQ	GEdit-Bench-EN Overall
Step1X-Edit (Liu et al., 2025)	7B+12B	3.06	–	–	–
BAGEL (Deng et al., 2025)	7B+7B	3.20	7.36	6.83	6.52
UniWorld-V1 (Lin et al., 2025a)	7B+12B	3.26	–	–	–
OmniGen2 (Wu et al., 2026)	3B+4B	3.44	7.16	6.77	6.41
UniLIP-1B (Tang et al., 2026)	1B+0.6B	3.81	–	–	–
InternVL-U (CoT) (Tian et al., 2026)	2B+1.7B	3.82	–	–	–
Step1X-Edit-v1.1 (Yin et al., 2026)	7B+12B	3.90	7.66	7.35	6.97
UniLIP-3B (Tang et al., 2026)	2B+1.6B	3.94	–	–	–
Ovis-U1 (Wang et al., 2025a)	2.4B+1.2B	4.00	–	–	6.42
MammothModa2 (Shen et al., 2025a)	8B+3B+2B	4.06	–	–	6.60
DeepGen 1.0 (Wang et al., 2026a)	3B+2B	4.14	–	–	7.17
LayerBridge-0.8B	0.8B+0.3B	3.86	7.23	6.95	6.82
LayerBridge-4B	4B+1.1B	4.16	7.82	7.34	7.28

Table 2: **Image editing.** Sizes follow Table 1, and Table 10 adds larger dedicated and proprietary editors. SC and PQ denote GEdit-Bench-EN semantic consistency and perceptual quality.

Configuration	GenEval	WISE	ImgEdit
State and time query (LayerBridge)	**0.87**	**0.54**	**3.86**
Final LLM outputs	0.77 (−0.10)	0.45 (−0.09)	3.28 (−0.58)
+ noisy-pixel input to the decoder	0.85 (−0.02)	0.46 (−0.08)	3.49 (−0.37)
One fixed layer per source	0.84 (−0.03)	0.48 (−0.06)	3.51 (−0.35)
Static per-site mixture	0.84 (−0.03)	0.49 (−0.05)	3.57 (−0.29)
LAP-style pooling, shared (Li et al., 2025)	0.85 (−0.02)	0.50 (−0.04)	3.69 (−0.17)
LAP-style pooling, per read site	0.85 (−0.02)	0.50 (−0.04)	3.71 (−0.15)
State-only query	0.85 (−0.02)	0.51 (−0.03)	3.74 (−0.12)
Time-only learned query	0.86 (−0.01)	0.49 (−0.05)	3.72 (−0.14)

Table 3: **Interface ablations at 0.8B.** Parentheses give differences from LayerBridge. The first two variants omit depth reads, the second adding $\mathbf{z}_t$ to the decoder input. The fixed-layer variant keeps one layer per source, and the others read the full bank (Section F).

(Table 12), versus 0.54/0.62 for LayerBridge. GenEval thus largely reflects the pretrained decoder, whereas VLM guidance raises WISE, mostly through depth reads (Table 3).

Image editing. At 0.8B/4B, LayerBridge reaches 3.86/4.16 on ImgEdit with fewer generation-specific parameters than UniLIP-1B/3B at 3.81/3.94 (Table 2). On GEdit-Bench-EN, our 4B model reaches 7.28, above DeepGen 1.0 at 7.17 and Step1X-Edit-v1.1 at 6.97.

Visual understanding. Over native Qwen3.5 (Table 1), MME-P rises from 1306.8 to 1327.5 at 0.8B and from 1569.1 to 1586.3 at 4B, while MMBench and MMMU stay within 0.5 points. Gains concentrate on vision-centric CV-Bench and MMVP, while HRBench-4K, POPE and OCRBench stay within 0.4 points. With the same modules and training at 0.8B, depth access adds 15.9 MME-P and 2.0 MMVP points over a Final-ViT control (Table 7), and still adds 14.8 MME-P points when only understanding loss trains these reads (Table 18).

4.3 Ablation study

Intermediate access. To distinguish intermediate access from initialization and added capacity, we test depth reads with random and pretrained decoder weights. With random weights, reads raise GenEval from 0.61 to 0.78 (Table 12). With pretrained weights, a parameter-matched final-layer control scores 0.78 versus 0.87 for depth reads (Table 14). Feeding noisy pixels directly to the final-layer decoder recovers GenEval (0.85) but not WISE or ImgEdit (0.46/3.49 versus 0.54/3.86), and the same input brings no gain to LayerBridge. Spatially shuffling its vision bank at test time reduces GenEval to 0.77, so the trained model relies on position-aligned reads (Table 14).

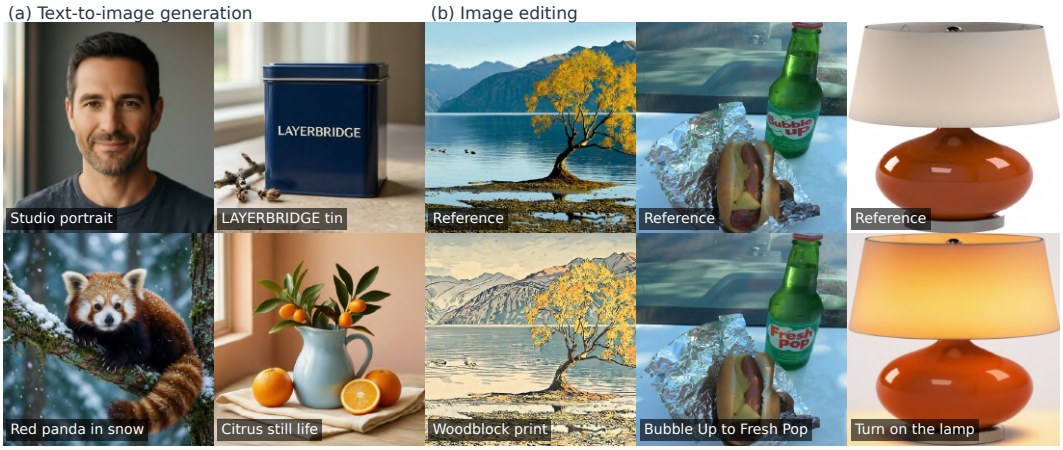


Figure 4: **Generation and editing with LayerBridge.** Generated examples span portraits, text rendering and natural scenes. Edits cover style, text and lighting, with references above.

State and time. With the same bank and per-block reads, removing time or state from the query lowers ImgEdit to 3.74/3.72 and WISE to 0.51/0.49, from 3.86/0.54 with both (Table 3). On this bank, static mixing reaches 3.57 and shared/per-site UniFusion-style layerwise attention pooling (LAP) 3.69/3.71. GenEval stays within 0.03 across these rules, as the pretrained decoder already handles composition, whereas the WISE and ImgEdit gaps have paired bootstrap intervals that exclude zero and recur at 4B (Table 17). Fixing the decoder mixture only early in denoising (Section F.5) lowers GenEval position accuracy from 0.82 to 0.75, whereas fixing it late raises unchanged-region LPIPS from 0.090 to 0.123, so a fixed mixture hurts layout early and detail late.

Noisy-token input. Token interpolation both attenuates noisy visual tokens entering the frozen LLM and adds a timestep embedding. To separate these roles, we keep the time term $(1-t)\psi(t)$ but pass $\mathbf{s}_t$ at full strength. This variant reaches 0.85 GenEval and 3.75 ImgEdit, between raw tokens (0.84/3.70) and full interpolation (0.87/3.86). Both the time term and suppressing noise-dominated tokens before they reach the LLM thus contribute to the gain (Table 15).

Depth-attention cost. Each read attends only over the layers at one image position. On one H800 GPU with batches of 16 images, the amortized sampling time per 512^2 image is 2.33/3.78 s at 0.8B/4B with depth reads and 1.92/2.98 s without them (Section B), while final-layer conditioning loses 0.58 ImgEdit at 0.8B (Table 3). Retaining all vision layers (20 bank entries instead of 12) leaves ImgEdit at 3.85 versus 3.86 for the default bank (Table 16).

Shared routes. Disabling reference reads at test time lowers ImgEdit from 3.86 to 3.20, or to 3.61/3.52 with only LLM or connector reads disabled (Table 14), and retraining without them still gives only 3.54. On 120 local edits, disabling them raises unchanged-region LPIPS from 0.090 to 0.171 and lowers instruction success from 79.2% to 71.7% (Table 19). Repeating the deepest LLM values reduces WISE from 0.54 to 0.48, and stating the required facts narrows this gap to 0.03 without closing it, consistent with intermediate LLM layers helping the decoder use knowledge that prompts leave implicit (Section F.7).

5 Conclusion

Understanding, generation and editing place different demands on visual detail and semantic context. LayerBridge accommodates these differences by letting each task select features from a shared frozen VLM hierarchy. The resulting models use compact generation modules at both backbone scales while supporting competitive generation and editing alongside vision-centric perception gains. These results suggest that adaptive access can reconcile task-specific information needs with a shared pretrained backbone, allowing unified multimodal models to extend from visual understanding to image creation while drawing on the same frozen hierarchy.

Reproducibility statement

Sections A and 3 specify the architecture, layer indices, depth-read pseudocode (Algorithm 1), initialization, freezing policy, loss functions, optimizer and stage composition. Section B documents preprocessing, sampling, benchmark splits, scoring and aggregation. We will release code and model weights. The probe protocols and controlled comparisons, including calibration sets, paired inputs and intervention definitions, appear in Sections E and F.

Ethics statement

Image generation and editing models can be misused to fabricate or manipulate images of real people and events. The frozen VLM and its training data can encode social biases, which generated and edited images may reproduce. Our training sources include public caption corpora and synthetic images distilled from proprietary generators. These sources have separate terms governing data use.

AI use statement

We used large language models to aid and polish the writing of this paper. LLM-based assistants also took part in discussions of experiments. The authors verified all methods, results and conclusions and take full responsibility for the paper.

References

Xiang An et al. LLaVA-OneVision-1.5: Fully Open Framework for Democratized Multimodal Training. *arXiv preprint arXiv:2509.23661*, 2025.

Black Forest Labs. FLUX.1. Official model release, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, et al. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv preprint arXiv:2506.15742*, 2025.

Qi Cai, Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. HiDream-I1: An open-source high-efficient image generative foundation model. In *ACM MM*, 2025.

Chameleon Team. Chameleon: Mixed-Modal Early-Fusion Foundation Models. *arXiv preprint arXiv:2405.09818*, 2024.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, et al. BLIP3-o: A Family of Fully Open Unified Multimodal Models-Architecture, Training and Dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Jiuhai Chen, Le Xue, Zhiyang Xu, Xichen Pan, Shusheng Yang, Can Qin, et al. BLIP3o-NEXT: Next frontier of native image generation. *arXiv preprint arXiv:2510.15857*, 2025b.

Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, et al. ShareGPT-4o-Image: Aligning Multimodal Models with GPT-4o-Level Image Generation. *arXiv preprint arXiv:2506.18095*, 2025c.

Shoufa Chen, Chongjian Ge, Shilong Zhang, Peize Sun, and Ping Luo. PixelFlow: Pixel-Space Generative Models with Flow. *arXiv preprint arXiv:2504.07963*, 2025d.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, et al. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. *arXiv preprint arXiv:2501.17811*, 2025e.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, et al. Emerging Properties in Unified Multimodal Pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, Yan Li, Yubo Wang, Zhijie Cao, Zhiqian Lin, Zhitao Yang, Zhongang Cai, Yuwei Niu, Yue Zhu, Bo Liu, Chengguang Lv, et al. SenseNova-U1: Unifying Multimodal Understanding and Generation with NEO-unify Architecture. *arXiv preprint arXiv:2605.12500*, 2026.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, et al. DreamLLM: Synergistic Multimodal Comprehension and Creation. In *ICLR*, 2024.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, et al. VLMEvalKit: An Open-Source Toolkit for Evaluating Large Multi-Modality Models. In *ACM MM*, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *ICML*, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, et al. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. In *NeurIPS Datasets and Benchmarks Track*, 2025.

Fengyi Fu, Mengqi Huang, Shaojin Wu, Yunsheng Jiang, Yufei Huo, Hao Li, et al. Lance: Unified Multimodal Modeling by Multi-Task Synergy. *arXiv preprint arXiv:2605.18678*, 2026.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, et al. BLINK: Multimodal Large Language Models Can See but Not Perceive. In *ECCV*, 2024.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment. In *NeurIPS Datasets and Benchmarks Track*, 2023.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *JAIR*, 47:853–899, 2013. doi: 10.1613/jair.3994.

Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, et al. Simpler Diffusion: 1.5 FID on ImageNet512 with Pixel-space Diffusion. In *CVPR*, 2025.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip Diffusion Models with LLM for Enhanced Semantic Alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *CVPR*, 2017.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. Online optimizer reference implementation, 2024.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*, pp. 787–798, 2014.

Kimi Team, Guangyu Chen, Yu Zhang, Jianlin Su, Weixin Xu, Siyuan Pan, Yaoyu Wang, Yucheng Wang, Guanduo Chen, Bohong Yin, Yutian Chen, Junjie Yan, Ming Wei, Y. Zhang, Fanqing Meng, Chao Hong, Xiaotong Xie, Shaowei Liu, Enzhe Lu, Yunpeng Tai, et al. Attention Residuals. *arXiv preprint arXiv:2603.15031*, 2026.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *EMNLP*, 2004.

Bozhou Li, Yushuo Guan, Haolin Li, Bohan Zeng, Yiyan Ji, Yue Ding, Pengfei Wan, Kun Gai, Yuanxing Zhang, and Wentao Zhang. Semantic Routing: Exploring Multi-Layer LLM Feature Weighting for Diffusion Transformers. *arXiv preprint arXiv:2602.03510*, 2026a.

Fan Li, Chonghuinan Wang, Lina Lei, Yuping Qiu, Jiaqi Xu, Jiaxiu Jiang, et al. HP-Edit: A Human-Preference Post-Training Framework for Image Editing. In *CVPR*, 2026b.

Kevin Li, Manuel Brack, Sudeep Katakol, Hareesh Ravi, and Ajinkya Kale. UniFusion: Vision-Language Model as Unified Encoder in Image Generation. *arXiv preprint arXiv:2510.12789*, 2025.

Shufan Li, Jiuxiang Gu, Kangning Liu, Zhe Lin, Zijun Wei, Aditya Grover, and Jason Kuen. Lavida-O: Elastic Large Masked Diffusion Models for Unified Multimodal Understanding and Generation. In *ICLR*, 2026c.

Tianhong Li and Kaiming He. Back to Basics: Let Denoising Generative Models Denoise. In *CVPR*, 2026.

Xu Li, Yi Zheng, Haotian Chen, Xiaolei Chen, Yuxuan Liang, Chenghang Lai, et al. Instruction-Guided Fusion of Multi-Layer Visual Features in Large Vision-Language Models. *Pattern Recognition*, 170:111932, 2026d.

Yanghao Li, Rui Qian, Bowen Pan, Haotian Zhang, Haoshuo Huang, Bowen Zhang, et al. MANZANO: A Simple and Scalable Unified Multimodal Model with a Hybrid Vision Tokenizer. In *ICLR*, 2026e.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *EMNLP*, 2023.

Bin Lin, Zongjian Li, Xinhua Cheng, et al. UniWorld-V1: High-Resolution Semantic Encoders for Unified Visual Understanding and Generation. *arXiv preprint arXiv:2506.03147*, 2025a.

Han Lin, Jaemin Cho, Amir Zadeh, Chuan Li, and Mohit Bansal. Bifrost-1: Bridging multimodal LLMs and diffusion models with patch-level CLIP latents. In *NeurIPS*, 2025b.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. In *CVPR*, 2017.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *ICLR*, 2023.

Bingchen Liu et al. Playground v3: Improving Text-to-Image Alignment with Deep-Fusion Large Language Models. *arXiv preprint arXiv:2409.10695*, 2024a.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, et al. Step1X-Edit: A Practical Framework for General Image Editing. *arXiv preprint arXiv:2504.17761*, 2025.

Xiangyue Liu, Zijian Zhang, Miles Yang, Zhao Zhong, Liefeng Bo, and Ping Tan. Rosetta: Composable Native Multimodal Pretraining. *arXiv preprint arXiv:2607.00293*, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *ICLR*, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, et al. MMBench: Is Your Multi-modal Model an All-around Player? In *ECCV*, 2024b.

Yuliang Liu, Zhang Li, Mingxin Huang, et al. OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models. *Science China Information Sciences*, 67(12), 2024c.

Xu Ma, Yitian Zhang, Qihua Dong, and Yun Fu. Fine-T2I: An Open, Large-Scale, and Diverse Dataset for High-Quality T2I Fine-Tuning. *arXiv preprint arXiv:2602.09439*, 2026a.

Zehong Ma, Longhui Wei, Shuai Wang, et al. DeCo: Frequency-Decoupled Pixel Diffusion for End-to-End Image Generation. In *CVPR*, 2026b.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, et al. LongCat-Image Technical Report. *arXiv preprint arXiv:2512.07584*, 2025.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. DeepStack: Deeply Stacking Visual Tokens is Surprisingly Simple and Effective for LMMs. In *NeurIPS*, 2024.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, et al. WISE: World Knowledge-Informed Semantic Evaluation for Text-to-Image Generation. In *ICML*, 2026.

Matteo Pagliardini, Amirkeivan Mohtashami, François Fleuret, and Martin Jaggi. DenseFormer: Enhancing information flow in transformers via depth weighted averaging. In *NeurIPS*, 2024.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, et al. Transfer between Modalities with MetaQueries. *arXiv preprint arXiv:2504.06256*, 2025.

Xichen Pan, Aashu Singh, Satya Narayan Shukla, Xiangjun Fan, Shlok Kumar Mishra, and Saining Xie. RepFusion: Leveraging Multimodal Priors for Denoising in Representation Space. *arXiv preprint arXiv:2606.14700*, 2026.

PKU-YuanGroup. WISE_Verified: Updated Evaluation Protocol and Leaderboard. Official benchmark code and leaderboard, 2026.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *ICLR*, 2024.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, et al. TokenFlow: Unified Image Tokenizer for Multimodal Understanding and Generation. In *CVPR*, 2025.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents. Qwen3.5 official model release announcement, February 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022.

Tao Shen, Xin Wan, Taicai Chen, Rui Zhang, Junwen Pan, Dawei Lu, Fanding Lei, Zhilin Lu, Yunfei Yang, Chen Cheng, Qi She, Chang Liu, and Zhenbang Sun. MammothModa2: A Unified AR-Diffusion Framework for Multimodal Understanding and Generation. *arXiv preprint arXiv:2511.18262*, 2025a.

Ying Shen, Zhiyang Xu, Jiuhai Chen, Shizhe Diao, Jiaxin Zhang, Yuguang Yao, et al. LaTtE-Flow: Layerwise Timestep-Expert Flow-based Transformer. *arXiv preprint arXiv:2506.06952*, 2025b.

Ying Shen, Tianrong Chen, Yuan Gao, Yizhe Zhang, Yuyang Wang, Miguel Ángel Bautista, et al. STARFlow2: Bridging Language Models and Normalizing Flows for Unified Multimodal Generation. *arXiv preprint arXiv:2605.08029*, 2026.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. LMFusion: Adapting Pretrained Language Models for Multimodal Generation. In *NeurIPS*, 2025.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved Baselines with Representation Autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

Hao Tang, Chenwei Xie, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. UniLiP: Adapting CLIP for Unified Multimodal Understanding, Generation and Editing. In *ICLR*, 2026.

Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, et al. InternVL-U: Democratizing Unified Multimodal Models for Understanding, Reasoning, Generation and Editing. *arXiv preprint arXiv:2603.09877*, 2026.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, et al. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. In *NeurIPS*, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In *CVPR*, 2024b.

Dianyi Wang, Ruihang Li, Feng Han, Chaofan Ma, Wei Song, Siyuan Wang, et al. DeepGen 1.0: A Lightweight Unified Multimodal Model for Advancing Image Generation and Editing. *arXiv preprint arXiv:2602.12205*, 2026a.

Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, et al. Ovis-U1 Technical Report. *arXiv preprint arXiv:2506.23044*, 2025a.

Shuai Wang, Ziteng Gao, Chenhui Zhu, et al. PixNerd: Pixel Neural Field Diffusion. In *ICLR*, 2026b.

Shuai Wang, Liang Li, Yang Chen, et al. UniDDT: Unifying Multimodal Understanding and Generation with Decoupled Diffusion Transformer. In *ECCV*, 2026c.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, and Dacheng Tao. Divide, Conquer and Combine: A Training-Free Framework for High-Resolution Image Perception in Multimodal Large Language Models. In *AAAI*, 2025b.

Xianbang Wang, Hanhong Zhao, Yiyang Lu, Kangyang Zhou, Linrui Ma, and Kaiming He. MiniT2I: A Minimalist Baseline for Text-to-Image Generation. Code repository on GitHub, 2026d.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, et al. Multimodal learning with next-token prediction for large multimodal models. *Nature*, 650:327–333, 2026e.

Yuhan Wang, Siwei Yang, Bingchen Zhao, Letian Zhang, Qing Liu, Yuyin Zhou, and Cihang Xie. GPT-IMAGE-EDIT-1.5M: A Million-Scale, GPT-Generated Image Dataset. *arXiv preprint arXiv:2507.21033*, 2025c.

Xinyu Wei, Guoli Yang, Jialu Zhou, Mingyue Yang, Leqian Li, Kedi Zhang, and Chunping Qiu. Dynamic Embedding of Hierarchical Visual Features for Efficient Vision-Language Fine-Tuning. *arXiv preprint arXiv:2508.17638*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, et al. Qwen-Image Technical Report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, et al. Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation. In *CVPR*, 2025b.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, et al. OmniGen2: Towards Instruction-Aligned Multimodal Generation. In *CVPR*, 2026.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, et al. VILA-U: a Unified Foundation Model Integrating Visual Understanding and Generation. In *ICLR*, 2025c.

Da Xiao, Qingye Meng, Shengping Li, and Xingyuan Yuan. MUDDFormer: Breaking residual bottlenecks in transformers via multiway dynamic dense connections. In *ICML*, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, et al. Show-o: One Single Transformer to Unify Multimodal Understanding and Generation. In *ICLR*, 2025a.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved Native Unified Multimodal Models. In *NeurIPS*, 2025b.

Chao Xu, Maohua Li, Qirui Li, Yixuan Xu, Yanke Zhou, Yunhe Li, et al. Rethinking Cross-Layer Information Routing in Diffusion Transformers. *arXiv preprint arXiv:2605.20708*, 2026.

Junzhe Xu, Yuyang Yin, and Xi Chen. TBAC-UniImage: Unified Understanding and Generation by Ladder-Side Diffusion Tuning. *arXiv preprint arXiv:2508.08098*, 2025.

Jinbo Yan, Limeng Qiao, Jie Qin, Junyan He, Feize Wu, and Guanglu Wan. UniSpace: Unified Visual Representation and Scalable Multimodal Modeling. *arXiv preprint arXiv:2608.08676*, 2026.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. MMaDA: Multimodal Large Diffusion Language Models. In *NeurIPS*, 2025a.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated Delta Networks: Improving Mamba2 with Delta Rule. In *ICLR*, 2025b.

Huanjin Yao, Wenhao Wu, Taojiannan Yang, YuXin Song, Mengxi Zhang, Haocheng Feng, et al. Dense Connector for MLLMs. In *NeurIPS*, 2024.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, et al. ImgEdit: A Unified Image Editing Dataset and Benchmark. In *NeurIPS Datasets and Benchmarks Track*, 2025.

Fukun Yin, Shiyu Liu, Yucheng Han, Zhibo Wang, Peng Xing, Rui Wang, et al. ReasonEdit: Towards Reasoning-Enhanced Image Editing Models. In *CVPR*, 2026.

Yongsheng Yu, Wei Xiong, Weili Nie, et al. PixelDiT: Pixel Diffusion Transformers for Image Generation. In *CVPR*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, et al. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *CVPR*, 2024.

Boyi Zeng, Yiqin Hao, Zitong Wang, Shixiang Song, He Li, Feichen Song, et al. Depth-Attention: Cross-Layer Value Mixing for Language Models. *arXiv preprint arXiv:2606.05014*, 2026.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 2018.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion Transformers with Representation Autoencoders. In *ICLR*, 2026.

Chunting Zhou, Lili Yu, Arun Babu, et al. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. In *ICLR*, 2025.

Defa Zhu, Hongzhi Huang, Zihao Huang, Yutao Zeng, Yunyao Mao, Banggu Wu, Qiyang Min, and Xun Zhou. Hyper-connections. In *ICLR*, 2025.

Lianghui Zhu, Yuxin Fang, Bencheng Liao, Shijie Wang, Tianheng Cheng, Zilong Huang, et al. Mixture-of-Depths Attention. *arXiv preprint arXiv:2603.15619*, 2026a.

Xuanyu Zhu, Yan Bai, Yang Shi, Yihang Lou, Yuanxing Zhang, Jing Jin, and Yuan Zhou. Beyond the Last Layer: Multi-Layer Representation Fusion for Visual Tokenization. *arXiv preprint arXiv:2605.10780*, 2026b.

APPENDIX ROADMAP

Sections A and B specify the architecture, training and evaluation protocols. Sections C and D report complete task results, extended comparisons and parameter scopes. Section E presents the frozen-feature probes, and Section F reports interface controls and extended ablations. Section G discusses the remaining limitations.

A HYPERPARAMETERS AND DATA

Understanding data. We sample 1M image-text examples from LLaVA-OneVision-1.5-Instruct (An et al., 2025) for Stages 1–2 and retain 50k for Stage 3. General instruction following, document/chart reading, spatial reasoning and visual knowledge/reasoning account for 30%, 30%, 10% and 30%, respectively, at every stage. Both scales use the same examples with seed 42, dynamic image resolution and cross-entropy on assistant answers.

Task balance and schedule. Understanding receives 25% of optimizer updates. The other updates use generation in Stage 1 and generation/editing at a 2:1 ratio in Stages 2–3. Each update uses one task batch. Pools are shuffled independently and reshuffled when exhausted. Updates and base learning rates per stage appear in Table 5. Each stage uses 1% linear warmup followed by cosine decay to 10% of its base learning rate.

Joint-model configuration	0.8B	4B
Vision / language layers	12 / 24	24 / 32
Visual taps (zero-indexed)	2, 5, 8, 11	2, 5, . . . , 23
LLM taps (full-attention)	3, 7, . . . , 23	3, 7, . . . , 31
LLM blocks with depth reads	same as LLM taps	same
Resolution, patch / merged tokens	512^2, 1024 / 256	same
Connector	6 blocks, full sequence, taps 1, 5	same (wider)
Decoder blocks (read sites)	17 (1, 4, 7, 10, 13, 16)	same
Depth reads per read site	2 (attention and MLP)	same
Decoder width / patch size	768 / 16	same
LLM / connector width	1024	2560
Bank entries (ViT + LLM + connector)	$4 + 6 + 2 = 12$	$8 + 8 + 2 = 18$
Shared vision-to-LLM modules	0.1B	0.7B
Generation-specific connector	0.1B	0.6B
Other generation-specific modules	0.2B	0.5B
Generation-specific subtotal	0.3B	1.1B
All added modules	0.4B	1.8B
Optimizer	Muon + AdamW, stage-wise lr, wd 0, clip 1.0	same
Image-task global batch / accumulation	512 / 1	same
Logit-normal t mean / std	-0.8 / 0.8	same
Alignment weight / type	1 / normalized MSE	same
Sampler	30-step Heun, final Euler step, CFG 4.5	same

Table 4: **Joint-model architecture and parameter counts.** Decoder sites are one-based and other taps zero-based. The all-added count includes shared weights once, and model names denote nominal backbone scales rather than exact parameter counts.

Stage	Generation	Editing	Understanding	Updates	Base LR
S1	3M	–	1M	50k	10^{-4}
S2	3M	1.1M	1M	100k	10^{-4}
S3	100k	40k	50k	20k	2×10^{-5}

Table 5: **Training data and schedule at both scales.** Data sizes and optimizer updates are rounded. Generation/editing use global batch 512 and understanding uses 256. Stage 3 covers 50 passes through the generation pool, including BLIP3o-60k.

Layer sampling and read placement. We retain every third ViT block, all full-attention LLM blocks and connector blocks 2 and 6, with exact indices in Table 4. This schedule is fixed independently of the probes, and Section F.2 tests stride, offset and decoder depth.

Spatial alignment. Qwen3.5 merges each 2×2 group of vision patches into one LLM image token. Shared LayerNorm-MLP projections align vision features to this grid for LLM depth attention. Separate projections align vision features to the decoder patch grid, and affine sub-pixel upsamplers bring LLM features to the same grid. Connector features already lie on this grid.

Connector and initialization. The 6-block connector uses freshly initialized Qwen3.5 full-attention blocks. Before the connector, a sub-pixel upsampler expands final normalized LLM image states from 256 to 1024 patch-grid tokens with distinct positions, retaining the text order. Connector blocks 2 and 6 read vision features before attention and before the MLP. Its projected final target states provide the decoder input. Both scales initialize the decoder from the 17 image-stream blocks, timestep embedder and output head of MiniT2I-B/16.

The bank stores complete post-block states. Vision features precede the final merger, LLM features precede the final normalization, and connector features use no additional output normalization. Depth attention uses 64-dimensional heads and RMSNorm with $\epsilon = 10^{-6}$. Projections are specific to each depth-attention module. LLM and connector queries share the timestep embedding ψ, while decoder queries use a separate embedding at decoder width.

Algorithm 1 Cross-network depth reads in one block. Each image position reads the bank of its own image, with $\tau = 1$ for references and clean images and $\tau = t$ for the noisy target.

Require: states $\mathbf{u} \in \mathbb{R}^{N \times d}$, image positions $\mathcal{I}$ with aligned banks $\mathbf{H}_{\mathcal{I}}$ and timesteps $\tau_{\mathcal{I}}$, parameters $\theta_{\text{attn}}, \theta_{\text{mlp}}$
Ensure: updated states $\mathbf{u}'$

function DEPTHREAD($\mathbf{v}, \mathbf{H}, \tau; \theta$) ▷ $\mathbf{v}$: image-position states
  $\bar{\mathbf{v}} \leftarrow R(\mathbf{v}) + A_\tau(\mathbf{e}_\tau)$ ▷ state and time form the query input
  $\bar{\mathbf{H}} \leftarrow R(\mathbf{H})$
  for each head h **in parallel do**
    $\mathbf{q}_h \leftarrow R_q(A_{q,h}(\bar{\mathbf{v}}))$, $\mathbf{K}_h \leftarrow R_k(A_{k,h}(\bar{\mathbf{H}}))$, $\mathbf{V}_h \leftarrow A_{v,h}(\bar{\mathbf{H}})$
    $\mathbf{a}_h \leftarrow \text{softmax}_K\left(\mathbf{q}_h \mathbf{K}_h^\top / \sqrt{d_h}\right)$ ▷ per position, over K layers
  end for
  $\Delta \leftarrow A_o\big(\text{concat}_h[\mathbf{a}_h \mathbf{V}_h]\big)$
  return $(\mathbf{W}_g \Delta + \mathbf{b}_g) \odot \Delta$ ▷ zero-initialized gate
end function
$\mathbf{b} \leftarrow \mathbf{u}$; $\mathbf{b}_{\mathcal{I}} \leftarrow \mathbf{u}_{\mathcal{I}} + \text{DEPTHREAD}(\mathbf{u}_{\mathcal{I}}, \mathbf{H}_{\mathcal{I}}, \tau_{\mathcal{I}}; \theta_{\text{attn}})$ ▷ text positions unchanged
$\mathbf{u} \leftarrow \mathbf{u} + \text{Attn}(\text{LN}_1(\mathbf{b}))$ ▷ original sequence attention
$\mathbf{b} \leftarrow \mathbf{u}$; $\mathbf{b}_{\mathcal{I}} \leftarrow \mathbf{u}_{\mathcal{I}} + \text{DEPTHREAD}(\mathbf{u}_{\mathcal{I}}, \mathbf{H}_{\mathcal{I}}, \tau_{\mathcal{I}}; \theta_{\text{mlp}})$
$\mathbf{u}' \leftarrow \mathbf{u} + \text{MLP}(\text{LN}_2(\mathbf{b}))$ ▷ original MLP
return $\mathbf{u}'$

Parameter scope. Generation-specific counts include the connector, its depth reads, image projections, vision modulation and pixel decoder. They exclude the frozen VLM and shared vision-to-LLM modules. Decoder blocks, their depth reads and the pixel head total about 0.2B parameters at both scales. Table 4 reports the shared components and all added parameters.

Joint optimization. The pixel objective follows JiT (Li & He, 2026), with RGB targets scaled to $[-1, 1]$. In Equation (4), $w(t) = \max(1 - t, 0.05)^{-2}$, except that $|t - 1| \leq 10^{-6}$ receives unit weight. The sampler converts predictions to velocities as $\hat{\mathbf{v}} = (\hat{\mathbf{x}} - \mathbf{z}_t)/\max(1 - t, 0.05)$, and its last step uses Euler integration to avoid evaluating the clean endpoint.

Global batches are 512 for generation/editing and 256 for understanding, with understanding cross-entropy scaled by $256/512$. Training uses 64 H100 GPUs without gradient accumulation and takes about 2/5 days at 0.8B/4B. Muon (Jordan et al., 2024) updates 2-D weight matrices except embeddings, normalizations, gates, feature projections and the output head, with momentum 0.95 and 5 Newton–Schulz iterations. AdamW handles the rest with $\beta = (0.9, 0.95)$ and $\epsilon = 10^{-8}$. Weight decay is zero and gradients are clipped at norm 1.

Attention masks and clean inputs. Editing orders the instruction, reference and noisy target in one sequence. Full-attention blocks are bidirectional within the noisy target span and causal elsewhere. DeltaNet remains causal. Only target states enter the decoder bank. References and understanding images use clean encoder tokens with $\psi(1)$ in depth queries. Understanding bypasses the connector and decoder, and the native comparison additionally disables the depth reads while retaining the original VLM preprocessing and masks.

One model for all tasks. At each scale, generation, editing and understanding results come from the same final checkpoint, without task-specific fine-tuning. Disabling its added reads recovers the original Qwen3.5, which supplies the native understanding scores. Controlled comparisons reuse identical prompts, noise seeds and editing inputs across conditions.

B EVALUATION PROTOCOLS

Generation settings. LayerBridge uses 512^2 images, CFG 4.5 and the 30-step Heun sampler for all generation and editing benchmarks. Prompts use the prefix `generate an image` in the Qwen chat template. The prompt-free branch removes the instruction but retains the current target and, for editing, the reference. Training drops text with probability 0.1 and always retains editing references. Prompt tokenization, attention masks and position embeddings are computed once per sample and

Model	MME-P	MME-C	Total	ΔTotal	ΔCount	ΔPos.	ΔOCR	ΔLandm.	ΔArt.	ΔCeleb.
LayerBridge-0.8B	1327.5	387.5	1715.0	+23.2	+8.3	+6.7	−2.5	+7.5	+1.0	−2.1
LayerBridge-4B	1586.3	562.5	2148.8	+21.2	+6.7	+3.3	−5.0	+9.3	+2.5	−1.5

Table 6: **Understanding profile by model size.** MME-P and MME-C are perception and cognition sums. Deltas are bridged minus native under the MME protocol in Section B. Selected category deltas do not completely decompose ΔTotal. The full perception-category profile is in Figure 5.

cached across sampling steps. The vision encoder receives no text, so both guidance branches share one vision forward pass.

Sampling time. The added interface attends over K feature layers independently at each of P image positions. Its attention products scale as $O(PKd)$ at width d, avoiding attention over the full PK spatial-depth bank. Source sampling and read placement control this added work without changing the backbone forward passes (Section F.2). We time sampling on one H800 GPU in bfloat16 with batches of 16 generated images, the default sampler, batched guidance branches and compiled kernels, excluding compilation and warm-up. Batch wall-clock time divided by 16 gives 2.33/3.78 s per 512^2 image at 0.8B/4B, compared with 1.92/2.98 s when depth reads and bank projections are removed. For understanding, depth reads occur only during image prefill, and autoregressive decoding requires no additional reads.

Sampling budget and quality. The default 30-step Heun sampler takes 29 Heun steps and a final Euler step, giving 59 guided evaluations with 2 branches each. At 0.8B, a 30-step Euler sampler roughly halves sampling time at 0.86 GenEval, and a 10-step Euler sampler needs 0.17 of the default time at 0.81. Encoding the target only at the first step and reusing these features for later steps lowers GenEval from 0.87 to 0.74, so the gains rely on refreshing VLM features as the target develops.

Benchmark settings. We follow the released evaluation scripts, using raw GenEval prompts with 4 samples each, ImgEdit Basic with the official GPT-4o rubric, GEdit-Bench-EN with GPT-4.1/VIEScore, and DPG with the COCO-large English mPLUG evaluator. Baseline scores and judge differences are documented in Section D. WISE uses the 1,000-prompt Verified release (PKU-YuanGroup, 2026) and its Qwen3.5-35B-A3B judge, with 1 image per prompt at each model's native resolution. We score MiniT2I, BLIP3-o, UniDDT, Ovis-U1 and UniLIP from released checkpoints with this official judge, which applies one rubric to every row. MME pairs native and bridged paths under the same VLMEvalKit parser, preprocessing and greedy decoding. MMBench-EN-DEV-v1.1 and MMMU validation follow the Qwen3.5 model-card modes (Qwen Team, 2026). At 0.8B and for MME, decoding is non-thinking and greedy. At 4B, MMBench and MMMU use thinking mode (temperature 0.6, top-p 0.95, top-k 20) and score the answer after the reasoning block. Native and bridged runs share these settings, and our native runs reproduce the published scores.

C Understanding results by category

Beyond the perception measures, MMBench and MMMU provide broader comparisons with the native model. Our scores reach 68.5/47.3 at 0.8B and 89.6/77.4 at 4B, within 0.5 points of paired native runs under the model-card modes.

Table 7 compares final ViT features with depth access using identical modules, data and schedules. At 0.8B, depth attention adds 15.9 MME-P points over the final-layer control. Relative to native, CV-Bench, MMVP and BLINK improve by 1.1–2.7 points at both scales and exceed the 0.8B Final-ViT control. POPE favors native 0.8B, and HRBench-4K and OCRBench favor both native scales. MME OCR declines by 2.5/5.0 points at 0.8B/4B. Within MME, the spatial categories counting and position rise by +8.3/+6.7 at 0.8B and +6.7/+3.3 at 4B, in line with the spatial gains on CV-Bench. Because the Final-ViT control shares the modules and training of the full model, its gap separates the contribution of intermediate features and query-dependent selection from the added visual route, as detailed in Section F.6.

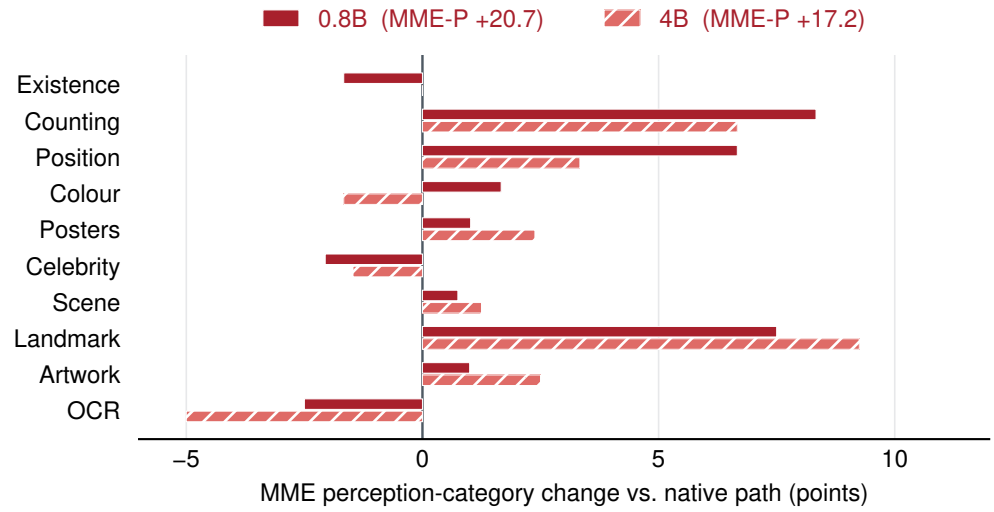


Figure 5: **MME perception-category profile.** Bars show bridged-minus-native differences for the same 0.8B and 4B models as Table 1. The native reference uses the frozen original VLM's MME evaluation. Unrounded perception-category scores sum to the displayed MME-P totals after rounding in Table 6. All 10 perception categories are shown. Cognition enters the total score separately.

Model	MME-P	CV-Bench	MMVP	BLINK	HRBench-4K	POPE	OCRBench
Qwen3.5-0.8B (native)	1306.8	64.2	26.0	47.5	65.4	87.7	76.8
Final ViT layer only	1311.6	64.4	26.7	47.5	64.6	87.4	76.4
LayerBridge-0.8B	1327.5	65.4	28.7	48.7	65.0	87.6	76.5
Qwen3.5-4B (native)	1569.1	80.3	60.7	52.6	75.6	87.0	85.3
LayerBridge-4B	1586.3	81.4	63.3	53.8	75.4	87.2	85.2

Table 7: **Fine-grained perception.** Paired VLMEvalKit evaluation on CV-Bench (Tong et al., 2024a), MMVP (Tong et al., 2024b), BLINK (Fu et al., 2024), HRBench-4K (Wang et al., 2025b), POPE (Li et al., 2023) and OCRBench (Liu et al., 2024c), with OCRBench divided by 10. Official aggregation is used, including CV-Bench weights $1/4, 1/4, 1/2$, paired MMVP accuracy, 4 HRBench answer rotations and pooled POPE F1. The 0.8B final-layer control repeats final ViT features across the bank with identical depth-attention modules and training.

D GENERATION AND EDITING RESULTS BY CATEGORY

Table 8 reports GenEval by category together with DPG-Bench and WISE. Table 9 reports all 9 ImgEdit categories. Parameter scopes are collected in Table 11.

Architecture and parameter comparisons. Tables 11 and 13 compare the conditioning interfaces and parameter components behind the task results.

SDXL GenEval follows Esser et al. (2024). The SD3-Medium/FLUX.1-dev GenEval and all three DPG scores follow Wu et al. (2025a), with denoiser counts from the model releases (Podell et al., 2024; Esser et al., 2024; Black Forest Labs, 2024). The remaining configurations follow their model reports (Chen et al., 2025a; Tang et al., 2026; Wang et al., 2026c; Xu et al., 2025; Shen et al., 2025a; Wang et al., 2026a; Li et al., 2026c) and Table 11. WISE uses official Verified scores for Janus-Pro, BAGEL and DeepGen and the same protocol for other models. BAGEL uses its no-CoT WISE score and the DPG score reported by Lance (Fu et al., 2026).

ImgEdit baselines follow UniLIP (Tang et al., 2026), Qwen-Image (Wu et al., 2025a) and the corresponding model reports (Wang et al., 2025a; Tian et al., 2026; Wang et al., 2026a; Meituan LongCat Team et al., 2025). Published category vectors and Overall scores are retained. Qwen-Image reports GPT-4.1 judging, while our ImgEdit protocol uses GPT-4o. GEdit scores follow those reports, HP-Edit for OmniGen2 (Li et al., 2026b) and ReasonEdit for Step1X-Edit-v1.1 (Yin et al., 2026).

Model	Single	Two	Count	Colors	Pos.	Attr.	GenEval	DPG	WISE
Janus-Pro-1B (Chen et al., 2025e)	0.98	0.82	0.51	0.89	0.65	0.56	0.73	82.6	0.27
Janus-Pro-7B (Chen et al., 2025e)	0.99	0.89	0.59	0.90	0.79	0.66	0.80	84.2	0.33
BAGEL (Deng et al., 2025)	0.99	0.94	0.81	0.88	0.64	0.63	0.82	85.1	0.35
OmniGen2 (Wu et al., 2026)	–	–	–	–	–	–	0.80	83.6	–
Show-o2-7B (Xie et al., 2025b)	–	–	–	–	–	–	0.76†	86.1	–
BLIP3-o-8B (Chen et al., 2025a)	–	–	–	–	–	–	0.84	81.6	0.52
MetaQuery-B (Pan et al., 2025)	–	–	–	–	–	–	0.74†	80.0	–
MetaQuery-L (Pan et al., 2025)	–	–	–	–	–	–	0.78†	81.1	–
MetaQuery-XL (Pan et al., 2025)	–	–	–	–	–	–	0.80†	82.1	–
RepFusion (pretrain) (Pan et al., 2026)	–	–	–	–	–	–	0.73	82.8	–
RepFusion-SFT (Pan et al., 2026)	–	–	–	–	–	–	0.87	85.1	–
Lavida-O (Li et al., 2026c)	0.99	0.85	0.71	0.86	0.65	0.58	0.77	81.8	–
Lavida-O + Planning (Li et al., 2026c)	0.99	0.94	0.75	0.90	0.84	0.68	0.85	82.9	–
Lavida-O + Reflection (Li et al., 2026c)	1.00	0.95	0.85	0.90	0.89	0.74	0.89	83.2	–
TBAC-UniImage (Xu et al., 2025)	0.99	0.94	0.77	0.92	0.83	0.75	0.87	81.0	–
MammothModa2 (Shen et al., 2025a)	1.00	0.97	0.63	0.89	0.90	0.82	0.87	87.2	–
UniDDT (Wang et al., 2026c)	0.99	0.93	0.71	0.92	0.85	0.80	0.87	86.9	0.58
UniLIP-1B (Tang et al., 2026)	0.99	0.92	0.83	0.91	0.83	0.80	0.88	–	0.48
UniLIP-3B (Tang et al., 2026)	1.00	0.95	0.84	0.91	0.86	0.84	0.90	–	0.54
LayerBridge-0.8B	0.98	0.93	0.80	0.90	0.82	0.79	0.87	84.7	0.54
LayerBridge-4B	0.99	0.95	0.85	0.91	0.87	0.83	0.90	88.2	0.62

Table 8: **Per-category GenEval.** Local evaluation uses raw prompts and 4 images per prompt. Baselines follow source settings. † marks rewritten prompts. WISE uses the official Verified scores for Janus-Pro and BAGEL (PKU-YuanGroup, 2026). BAGEL DPG is from Lance (Fu et al., 2026), and Janus-Pro-1B/7B categories follow Pan et al. (2025) and Chen et al. (2025a).

E LAYER-WISE PROBES OF THE FROZEN VLM

Figure 2 measures what different depths of frozen Qwen3.5-4B encode, using lightweight readouts trained on frozen features. Text precedes the image tokens, and full-attention blocks attend bidirectionally within the image, the mask that LayerBridge applies to target images during generation. Probe images are resized to 448^2, giving a 14×14 LLM grid. We read post-block image-token states without any LayerBridge module and compare depths under the same head, data and fitting budget within each task.

Referring-expression grounding. We use RefCOCO (Kazemzadeh et al., 2014) expressions on the original COCO images, split by source image into 2,000 fitting, 300 development and 1,000 test expressions. Each object mask is downsampled to the 14×14 LLM grid. At every probed layer, an identical MLP with 2 hidden layers predicts which grid cells belong to the referred object from 256-dimensional PCA features. Checkpoints are selected on the development set, and all scores below are test grid mIoU.

Using only expression-conditioned image-token states, L16 reaches 15.9 grid mIoU, versus 10.1 at L32 and 12.2 at V24. The paired 95% intervals of these gaps are 4.9–6.8 and 2.5–4.9 over 2,000 source-image bootstraps. Replacing the expression with one that refers to a different object lowers L16 by 5.00 points but L32 by only 1.55, so intermediate states are more prompt-specific.

Reconstruction and caption retrieval. Both probes use Flickr8k (Hodosh et al., 2013) and share 32 held-out test photos. For reconstruction, the same small spatial Transformer and residual CNN map features from V1, V24, L1, L16 and L32 to a 112^2 image, fitted on 256 photos. Held-out PSNR is 37.90/19.54 dB at V1/V24 and 14.52/13.53/12.61 dB at L1/L16/L32, so appearance detail concentrates in shallow vision layers.

For retrieval, the caption precedes the image, so image-token states are caption-conditioned as in generation. At each layer, a ridge head maps mean-pooled image features to the final-layer embedding of the photo's caption, and cosine similarity ranks the 32 test captions. Top-1 accuracy peaks at L12/L16 with 56.25%, versus 43.75% at L32, with a paired 95% interval of $[-3.1, 28.1]$ points for each gap over 1,000 bootstraps. Supplying the caption of a different photo lowers accuracy by up to

Model	Size	Add	Adj.	Extr.	Repl.	Rem.	Bg.	Style	Hyb.	Act.	Overall
GPT Image 1 [High]§	–	4.61	4.33	2.90	4.35	3.66	4.57	4.93	3.96	4.89	4.20
BAGEL	7B+7B	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20
OmniGen2	3B+4B	3.57	3.06	1.77	3.74	3.20	3.57	4.81	2.52	4.68	3.44
UniLIP-1B	1B+0.6B	4.11	3.58	1.94	4.30	3.97	4.00	4.87	3.17	4.35	3.81
UniLIP-3B	2B+1.6B	4.29	3.90	2.22	4.44	4.10	4.14	4.80	3.04	4.55	3.94
Ovis-U1	2.4B+1.2B	4.13	3.62	2.98	4.45	4.06	4.22	4.69	3.45	4.61	4.00
FLUX.1 Kontext [Pro]§	–	4.25	4.15	2.35	4.56	3.57	4.26	4.57	3.68	4.63	4.00
Qwen-Image-Edit§	7B+20B	4.38	4.16	3.43	4.66	4.14	4.38	4.81	3.82	4.69	4.27
LongCat-Image-Edit	7B+6B	4.51	4.57	3.93	4.76	4.60	4.49	4.85	4.01	4.74	4.50
LayerBridge-0.8B	0.8B+0.3B	4.20	3.60	2.05	4.32	4.12	4.14	4.80	3.18	4.33	3.86
LayerBridge-4B	4B+1.1B	4.60	4.05	2.42	4.60	4.45	4.45	4.82	3.35	4.68	4.16

Table 9: **Image editing on ImgEdit.** Scores range from 1 to 5. Baseline category vectors, Overall scores and component budgets follow the cited reports (Section D), so baseline Overall may differ from the mean of the rounded categories. LayerBridge Overall averages its 9 category scores equally. LayerBridge adds no editing-specific parameters. §ImgEdit scored with GPT-4.1 in the source report. Other rows use the official GPT-4o evaluator.

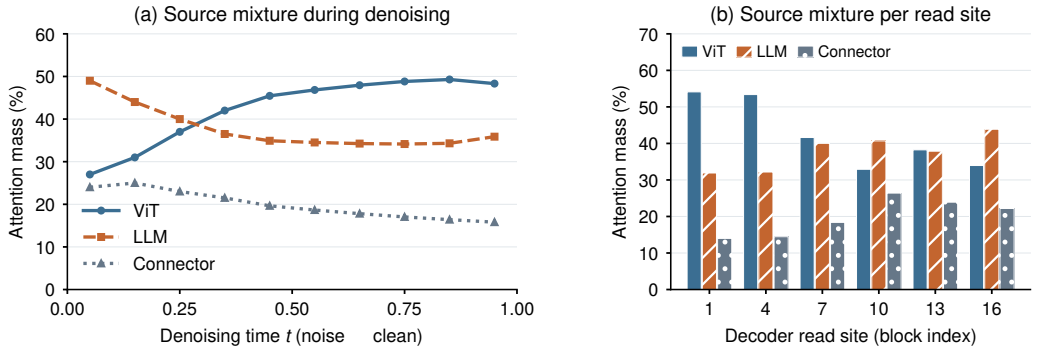


Figure 6: **Source-level depth usage of the decoder.** (a) Conditional-branch attention mass summed over all retained layers within each source, then averaged over 32 fixed prompts, read sites, heads and positions in 10 time bins. (b) Source mixtures at the 6 decoder blocks, averaging their attention and MLP reads over the same bins. All bank entries contribute to the three source totals, without separate normalization within a source.

21.9 points at L12, showing that intermediate states depend strongly on the supplied text. Section F tests the value of these depths for synthesis.

F Interface ablations and controls

Experiments in this section use LayerBridge-0.8B, except for the 4B comparisons in Table 17. The controls separate intermediate access from adaptive selection and test their roles in layout and detail preservation. Table 14 extends the comparisons in Table 3.

F.1 Decoder initialization and feature access

To isolate depth access from decoder pretraining, Table 12 compares final-only and full-depth conditions with randomly initialized decoders. Training data could also explain the gains over MiniT2I. We therefore train MiniT2I-B/16 with its own text conditioning on our generation data and schedule. It reaches 0.86 GenEval but only 0.42 WISE, whereas LayerBridge-0.8B reaches 0.87/0.54. After Stage 2, LayerBridge-0.8B already reaches 0.51 WISE, so the knowledge gain does not depend on the Stage 3 data. Later controls retain the pretrained decoder.

Model	Size	ImgEdit Overall	GEdit-Bench-EN SC	PQ	Overall
Step1X-Edit (Liu et al., 2025)	7B+12B	3.06	–	–	–
BAGEL (Deng et al., 2025)	7B+7B	3.20	7.36	6.83	6.52
TBAC-UniImage (Xu et al., 2025)	3B+1.6B	3.25	–	–	–
UniWorld-V1 (Lin et al., 2025a)	7B+12B	3.26	–	–	–
OmniGen2 (Wu et al., 2026)	3B+4B	3.44	7.16	6.77	6.41
InternVL-U (Tian et al., 2026)	2B+1.7B	3.67	–	–	–
Lavida-O (Li et al., 2026c)	8B+2.4B	3.71	–	–	–
Lavida-O + Planning (Li et al., 2026c)	8B+2.4B	3.80	–	–	–
UniLIP-1B (Tang et al., 2026)	1B+0.6B	3.81	–	–	–
InternVL-U (CoT) (Tian et al., 2026)	2B+1.7B	3.82	–	–	–
Step1X-Edit-v1.1 (Yin et al., 2026)	7B+12B	3.90	7.66	7.35	6.97
UniLIP-3B (Tang et al., 2026)	2B+1.6B	3.94	–	–	–
Ovis-U1 (Wang et al., 2025a)	2.4B+1.2B	4.00	–	–	6.42
FLUX.1 Kontext [Pro]§ (Black Forest Labs et al., 2025)	–	4.00	7.02	7.60	6.56
MammothModa2 (Shen et al., 2025a)	8B+3B+2B	4.06	–	–	6.60
DeepGen 1.0 (Wang et al., 2026a)	3B+2B	4.14	–	–	7.17
GPT Image 1 [High]§ (Wu et al., 2025a)	–	4.20	7.85	7.62	7.53
Qwen-Image-Edit§ (Wu et al., 2025a)	7B+20B	4.27	8.00	7.86	7.56
LongCat-Image-Edit (Meituan LongCat Team et al., 2025)	7B+6B	4.50	8.18	8.00	7.64
LayerBridge-0.8B	0.8B+0.3B	3.86	7.23	6.95	6.82
LayerBridge-4B	4B+1.1B	4.16	7.82	7.34	7.28

Table 10: **Extended image-editing comparison.** Sizes follow Table 1. Baseline scores follow published reports (Section D). ImgEdit Overall summarizes 9 edit types. GEdit-Bench-EN reports semantic consistency (SC), perceptual quality (PQ) and Overall. §ImgEdit scored with GPT-4.1 in the source report. Other rows use the official GPT-4o evaluator.

Connector capacity. Aligning token grids and widths does not itself adapt LLM states to pixel prediction. Table 15 therefore compares direct projection and upsampling with 2- and 6-block connectors. GenEval/ImgEdit rise from 0.80/3.30 without connector blocks to 0.85/3.70 with 2 and 0.87/3.86 with 6. The comparison supports learned adaptation beyond coordinate mapping, while changing connector capacity, read sites and exported taps together.

Component ablations. Noise adaptation must serve synthesis without changing the encoder's clean-image mapping. Removing the clean anchor tests this constraint and lowers MME-P by 16.9 points at unchanged GenEval (Table 15).

For noisy inputs, interpolation controls visual strength while alignment specifies the desired content. We separate these roles by retaining time information without attenuation, using raw tokens with alignment, or removing alignment while retaining interpolation. Raw input and no alignment reduce GenEval by 0.03 and 0.02. With the attenuation control, they support controlling both the strength and the content of visual evidence entering the LLM.

With the input design fixed, causal target attention tests whether each image position needs context from later positions. Bidirectional attention adds 0.02 GenEval, supporting within-image context during recovery. Gate and density variants test sensitivity to the integration design. Linear gates and sigmoid gates scaled by a zero-initialized vector perform similarly, while a sparser bank retains gains over final-layer conditioning (Tables 14 and 15). Section F.2 varies the read sites.

Final-layer system and capacity control. To test whether extra parameters alone explain the depth-interface gain, we add capacity without adding intermediate access. The *final-LLM-only system* retains the clean-anchored noisy encoder, connector and decoder initialization but disables all depth reads. The pixel-input controls embed $\mathbf{z}_t$ with the pretrained MiniT2I patch embedder and add it to the decoder input through a zero-initialized gate. The capacity control keeps the decoder width at 768 and loads the same 17 pretrained image-stream blocks. It adds a residual MLP after each of the 6 decoder read-site blocks, $\mathbf{u} \mapsto \mathbf{u} + W_2\,\mathrm{SiLU}(W_1\,\mathrm{RMSNorm}(\mathbf{u}))$, with W_2 initialized to zero. A common hidden width, rounded to a multiple of 64, matches the full model's active trainable parameter count within 0.5%. All 6 MLPs execute during training and sampling under the same

Model	Reported backbone	Generation-specific modules
BAGEL	7B (joint)	7B expert
Ovis-U1	2.4B (joint)	1.2B decoder
UniDDT	Qwen3-VL-4B (trained)	1B
UniLIP-1B / 3B	InternVL3-1B / 2B (partial)	0.6B / 1.6B DiT
BLIP3-o-8B	7B (frozen)	1.4B flow module
MetaQuery-B / L / XL	0.5B / 3B / 7B (frozen)	1.6B DiT
RepFusion-SFT	7B (frozen)	3.2B representation + 1.6B DiT
Lavida-O	8B understanding branch	2.4B generation branch
TBAC-UniImage	Qwen2.5-VL-3B (frozen)	1.6B DiT + 10M connector
MammothModa2	Qwen3-VL-8B (joint)	3B experts + 2B DiT
DeepGen 1.0	3B VLM	2B DiT
LongCat-Image-Edit	7B	6B diffusion editor
LayerBridge-0.8B / 4B	Qwen3.5-0.8B / 4B (frozen)	0.3B / 1.1B (0.4B / 1.8B with shared reads)

Table 11: **Parameter accounting for the compared systems.** Each row follows its source's module decomposition. "Partial" denotes partial backbone training. DiT-only counts exclude connectors and image decoders. Shared LayerBridge modules appear in Table 4.

Generation interface	GenEval	DPG	WISE	ImgEdit
Random-init decoder, final LLM outputs only	0.61	78.9	0.41	2.95
Random-init decoder, full depth reads	0.78	82.0	0.48	3.52
MiniT2I-B/16, our data and updates, no VLM	0.86	84.4	0.42	–
LayerBridge after Stage 2	0.84	84.1	0.51	3.72
LayerBridge, full model	0.87	84.7	0.54	3.86

Table 12: **Decoder initialization, data and stages.** Random-init rows use the same 17-block decoder and task exposure, with decoder weights trainable from Stage 1 in both controls. Their comparison tests depth access independently of decoder pretraining. The MiniT2I row trains the original text-conditioned generator on our generation data and schedule. The final row reports the complete model initialized from MiniT2I.

stage schedule and task exposure. This control reaches 0.78 GenEval, barely above 0.77 without the MLPs, so added capacity alone does not explain the 0.87 of depth reads. The pixel-input control instead tests direct access to the noisy target. It recovers GenEval but leaves WISE and ImgEdit 0.08 and 0.37 below the full model (Table 3), so depth reads retrieve more than pixels.

F.2 Source sampling and decoder placement

The layer schedule should expose complementary features without depending on a particular offset. The top panel of Table 16 varies source density and offset separately under the same training schedule. Shifting the taps by 1 or 2 layers removes the final ViT layer from the bank and costs up to 0.07 ImgEdit. Denser sampling changes ImgEdit by at most 0.02 but increases read cost. Stride $s = 3$ balances quality and readout cost without choosing task-specific layers from the probes.

The bottom panel varies the read sites within the pretrained 17-block decoder. Reducing them to 3 lowers ImgEdit by 0.10, while 9 sites add 0.01, so 6 sites capture nearly all of the benefit. Shifting the 6 sites by one block changes GenEval by at most 0.01. Deeper decoders stay within 0.02 of the default despite up to 42.5M more parameters, so we keep the full 17-block MiniT2I image stream.

F.3 Per-category interface ablations

An aggregate GenEval gain does not identify which visual relations benefit from depth access. We therefore decompose the route comparisons in Table 14 by category. Counting rises from 0.55 without depth reads to 0.72 with vision reads in the decoder, 0.75 with vision reads in the LLM and 0.80 for the full model, and position rises from 0.65 to 0.77, 0.78 and 0.82. Single-object prompts remain saturated, and attribute binding varies by at most 0.02. Depth reads thus mainly improve object counts and locations, complementing the grounding probes (Figure 2).

Method	Decoder conditions	Layer combination	Vision-to-LLM reuse
MetaQuery	Final learnable-query states	Final-layer interface	Native path
RepFusion	Noisy-RAE final states	Final-layer interface	Native path
UniFusion	Pooled prompt/reference layers	Source-feature attention	Native path
TBAC-UniImage	Layer-matched query states	Fixed layer pairing	Native path
MammothModa2	Pooled AR states	Uniform layer average	Native path
LayerBridge	ViT, LLM, connector	Decoder state and timestep per site	Shared depth reads

Table 13: **Conditioning-interface comparison.** Final-layer interfaces (Pan et al., 2025; 2026) condition on final states, whereas pooling and fixed layer assignments (Li et al., 2025; Xu et al., 2025; Shen et al., 2025a) mix conditioning-input layers. LayerBridge lets evolving decoder states set their depth mixtures. The LAP-style control in Table 3 uses our bank and training settings.

F.4 Matched decoder-bank and query comparisons

Giving the decoder more layers does not establish that its current state must determine their mixture. We separate access from selection by varying the bank first and the mixing rule with the full bank retained. The fixed-tap, pooling and query variants in Table 3 share upstream reads, the connector-to-decoder path, training, data order and evaluation seeds. A fixed deepest layer per source (V12, L24, C6) limits feature access at every read site. Static mixing retains the full bank and learns per-site, per-head weights shared across positions, images and times. LAP-style pooling applies UniFusion's rule to our target bank, with 2 self-attention blocks across projected depths at each position, followed by pooling and refinement of one condition for all 6 decoder read sites. The shared variant changes both the mixing rule and whether the mixture is shared across read sites. The per-site variant applies a separate pooling module to each of the 12 reads, two per read site, with matched total parameters, mirroring the per-layer variant that UniFusion reports as weaker than a shared condition (Li et al., 2025), so its gap to LayerBridge isolates the mixing rule.

The current state reflects image content, while the timestep explicitly identifies the noise stage. To test whether either cue alone suffices, we remove them separately while retaining the full bank. The time-only variant replaces $R(\mathbf{u}_p)$ in Equation (1) with a learned vector specific to each read site, retaining $A_\tau(\mathbf{e}_\tau)$. State-only removes this direct time term, retaining the state query. Keys, values, output projections, gates and later query operations are unchanged. Time-only queries still access image-dependent keys and values, while state-only queries can inherit noise information through the evolving state. Combining them gains 0.03/0.05 WISE and 0.12/0.14 ImgEdit over state-only/time-only queries, complementing the smaller GenEval gains.

Differences of a few hundredths could reflect which benchmark prompts happen to be sampled. Table 17 therefore reports paired bootstrap intervals (Koehn, 2004) with 10,000 resamples over GenEval prompts, WISE prompts and ImgEdit items, using identical noise seeds across variants. Each GenEval prompt is resampled together with its 4 images and the 6 category means are recomputed. Binary WISE scores are resampled within categories and recombined with the official weights, and ImgEdit items are resampled within each edit type before the equal-weight average. Every WISE and ImgEdit interval excludes zero, whereas GenEval intervals reach zero for most selection rules. At 4B, the pixel-input baseline, LAP-style pooling and state-only queries show the same ordering, so the advantage of state- and time-conditioned reads holds at both scales.

F.5 Denoising-window interventions

To test when adaptive selection matters, we fix decoder attention to a calibration-average depth distribution during selected denoising windows (Table 19). Separate calibration sets contain 128 generation prompts and 128 editing pairs outside the evaluation benchmarks. We average attention separately at each read site ℓ and head h over calibration examples, positions and evaluation times,

$$\pi_{\ell,h,k} = \frac{1}{NPT} \sum_{n=1}^{N} \sum_{p=1}^{P} \sum_{j=1}^{T} a_{n,\ell,p,h,k}(t_j), \tag{5}$$

where a is post-softmax attention, k indexes the full target bank, and T covers the model evaluations in the 30-step Heun sampler. Conditional and prompt-free branches use separately averaged distri-

Interface	GenEval	DPG	ImgEdit
What is read			
Final LLM outputs only, no depth reads	0.77	82.3	3.28
same interface, capacity-matched residual MLPs	0.78	82.7	3.31
+ vision hierarchy → decoder (depth attention)	0.84	83.9	3.45
+ vision hierarchy → LLM	0.85	84.1	3.71
+ LLM/connector hierarchy → decoder (LayerBridge)	0.87	84.7	3.86
Direct noisy-pixel input to the decoder			
Final LLM outputs + pixel input, no depth reads	0.85	83.8	3.49
LayerBridge + pixel input	0.86	85.0	3.84
How the same vision taps are read (vs. vision hierarchy → decoder)			
static sum over taps (RAEv2-style) (Singh et al., 2026)	0.80	83.1	3.37
static per-site mixture of vision taps	0.81	83.4	3.42
Interventions on the trained full model			
decoder query time permuted across steps	0.84	83.6	3.62
decoder reads fixed to the calibration-average distribution	0.83	83.2	3.55
vision bank spatially shuffled	0.77	81.5	3.21
reference reads disabled in the LLM	–	–	3.61
reference reads disabled in the connector	–	–	3.52
reference reads disabled in both	–	–	3.20
Retrained without reference depth reads			
reference reads removed in training and sampling	0.87	84.6	3.54

Table 14: **Feature-interface comparisons.** Successive rows add feature sources and routes. Pixel-input rows add a gated patch embedding of $\mathbf{z}_t$ to the decoder input. Static mixtures compare with the vision-to-decoder row. Test-time interventions keep the checkpoint fixed. Reference-read rows disable reads at reference positions. The other replacements affect decoder reads, and parameter budgets vary across controls.

butions. The trained checkpoint, values, projections and gates are unchanged and only the softmax weights are replaced, so the decoder cannot adjust its depth mixture within the intervened window.

Early replacement covers Heun steps 1–10 ($t = 0$ to $1/3$). Late replacement covers steps 20–29 ($t = 19/30$ to $29/30$). Each covers 20 guided evaluations, while query times and upstream computations keep their true values, and prompts, inputs and noise seeds are paired across conditions. Early layout is evaluated on the 100 GenEval position prompts with 4 samples each. Detail preservation is measured by unchanged-region LPIPS on the 120 local edits defined below.

To examine whether the selected sources change as image content emerges, Figure 6 tracks source-level attention on 32 fixed prompts. In the high-noise window ($t < 0.2$), LLM features receive 46.5% of the attention mass, compared with 29.0% for ViT features and 24.5% for the connector. Near the clean endpoint ($t > 0.8$), ViT mass rises to 48.8%, while LLM and connector mass become 35.1% and 16.1%. The window interventions in Table 19 show greater sensitivity of position accuracy to early replacement and preservation to late replacement. Because the full-trajectory average also changes the mixture at each noise level, these interventions do not separate content-dependent from time-dependent selection, which the query ablations in Table 3 test.

F.6 Shared-interface and reference-path comparisons

Adding a trained visual route can change understanding even without intermediate features. To distinguish this effect from depth access, we compare native Qwen3.5, final ViT layer only, and the full depth bank. Trained variants share protocol, data, schedule, generation path and parameter tensors. The final-layer control repeats V12 across the bank. Identical keys and values make its read independent of query and time, so equal parameter counts do not imply equal effective capacity.

The shared interface receives both understanding and image-task gradients, leaving supervision as another explanation for perception gains. Table 18 therefore crosses final-layer versus depth-bank features with understanding-only versus joint losses for vision-to-LLM depth attention. In the understanding-only condition, image-loss gradients are suppressed for the vision-to-LLM bank

Variant	GenEval	ImgEdit	MME-P
LayerBridge	0.87	3.86	1327.5
Unanchored vision AdaLN, $\mathrm{MLP}(\phi(t))$	0.87	3.83	1310.6
No token interpolation ($\tilde{\mathbf{s}}_t = \mathbf{s}_t$)	0.84	3.70	1326.2
No visual attenuation ($\tilde{\mathbf{s}}_t = \mathbf{s}_t + (1-t)\psi(t)$)	0.85	3.75	1327.1
No alignment loss ($\lambda = 0$)	0.85	3.76	1325.4
Causal attention inside the noisy target block	0.85	3.78	1328.1
Sigmoid gate in place of the linear gate	0.86	3.83	1326.0
Every second entry of both banks (decoder 6 of 12)	0.86	3.79	1322.8
Connector with 2 blocks instead of 6	0.85	3.70	–
No connector blocks (projection only)	0.80	3.30	–

Table 15: **Component ablations.** Each row changes one design choice and retrains under the same stages and exposure. The clean anchor mainly protects bridged understanding, whereas token interpolation, feature alignment and bidirectional target attention matter for synthesis. Connector variants keep 1 or 0 connector read sites and leave the understanding inference path unchanged.

Setting	s	r	Bank entries	GenEval	ImgEdit
Sparse	6	0	10	0.86	3.80
Default	3	0	12	0.87	3.86
Dense	2	0	14	0.86	3.88
All vision layers	1	0	20	0.87	3.85
Shift 1 layer	3	1	12	0.86	3.82
Shift 2 layers	3	2	12	0.86	3.79

Blocks	Read sites (one-based)	Added params.	GenEval	ImgEdit
17	4, 10, 16	−21.3M	0.85	3.76
17 (default)	1, 4, 7, 10, 13, 16	–	0.87	3.86
17	2, 5, 8, 11, 14, 17	–	0.86	3.83
17	1, 3, 5, …, 17	+21.3M	0.87	3.87
20	2, 5, 8, 12, 15, 19	+21.2M	0.86	3.88
23	2, 6, 10, 14, 18, 22	+42.5M	0.87	3.85

Table 16: **Source banks and decoder read sites.** Top, one-based vision taps are $D-r, D-r-s, \ldots > 0$ with vision depth $D = 12$, while LLM and connector taps, 17 decoder blocks and 6 read sites stay fixed. Bank counts include all three sources. Bottom, width is 768 throughout and added parameters are relative to the default. The 17-block rows keep the pretrained decoder, and the 20- and 23-block rows insert zero-initialized blocks between pretrained blocks.

projections and read modules, but not for ψ, vision modulation, connector and decoder, so shared modules still pass task information. The depth bank improves MME-P under both gradient settings.

F.7 Knowledge and reference-detail transfer

Intermediate LLM entries could help through knowledge or other context, so Table 19 changes both access to them and the facts in the prompt. In the language intervention, the values of all selected LLM entries in the decoder bank are replaced by the deepest selected LLM entry at the same target position. ViT and connector entries, decoder queries, attention probabilities and the final-LLM path are unchanged, so total attention mass on the language source is preserved.

For WISE, each original prompt has a paired version that explicitly supplies the factual entity, attribute or relation needed to depict the requested scene. Expansions follow the benchmark's knowledge specification without aesthetic instructions or model outputs, and both versions keep the original scoring requirement. The gap between full and deepest-only LLM entries falls from 0.06 on original prompts to 0.03 on fact-explicit prompts. Stating the facts narrows but does not close the gap, consistent with intermediate LLM layers supplying knowledge that prompts leave implicit.

Reference reads should support the requested change while preserving regions outside it. We therefore evaluate preservation and change success separately on 30 examples each of addition, removal,

Scale	Variant, Δ vs. LayerBridge	GenEval	WISE	ImgEdit
0.8B	Final layer + pixel input	$-0.02\,[-0.04, -0.01]$	$-0.08\,[-0.11, -0.05]$	$-0.37\,[-0.43, -0.31]$
	Static per-site mixture	$-0.03\,[-0.05, -0.01]$	$-0.05\,[-0.08, -0.02]$	$-0.29\,[-0.35, -0.23]$
	LAP-style pooling	$-0.02\,[-0.04, 0.00]$	$-0.04\,[-0.07, -0.02]$	$-0.17\,[-0.23, -0.11]$
	LAP-style, per read site	$-0.02\,[-0.04, 0.00]$	$-0.04\,[-0.06, -0.02]$	$-0.15\,[-0.21, -0.10]$
	State-only query	$-0.02\,[-0.03, 0.00]$	$-0.03\,[-0.05, -0.01]$	$-0.12\,[-0.18, -0.07]$
	Time-only query	$-0.01\,[-0.03, 0.01]$	$-0.05\,[-0.07, -0.03]$	$-0.14\,[-0.19, -0.08]$
4B	Final layer + pixel input	$-0.02\,[-0.03, 0.00]$	$-0.09\,[-0.12, -0.06]$	$-0.35\,[-0.41, -0.29]$
	LAP-style pooling	$-0.01\,[-0.03, 0.01]$	$-0.04\,[-0.06, -0.02]$	$-0.15\,[-0.20, -0.10]$
	State-only query	$-0.01\,[-0.02, 0.01]$	$-0.03\,[-0.05, -0.01]$	$-0.11\,[-0.16, -0.06]$

Table 17: **Robustness of the main interface comparisons.** Differences from LayerBridge at the same scale, with 95% paired bootstrap intervals over benchmark prompts or edit items. The 4B variants reach 0.88/0.53/3.81, 0.89/0.58/4.01 and 0.89/0.59/4.05 on GenEval/WISE/ImgEdit.

Vision-to-LLM read	Training losses	MME-P	ImgEdit
Final ViT layer only	Understanding only	1309.0	3.70
Depth bank	Understanding only	1323.8	3.79
Final ViT layer only	Joint	1311.6	3.74
Depth bank	Joint (LayerBridge)	1327.5	3.86

Table 18: **Depth features and training losses.** Understanding-only rows train the shared vision-to-LLM depth attention with the understanding loss alone (Section F.6).

Read condition	GenEval position	Unchanged LPIPS ↓
Adaptive throughout	0.82	0.090
Fixed in early window	0.75	0.095
Fixed in late window	0.81	0.123
Comparison	**Original prompts**	**Facts supplied**
Full model, WISE	0.54	0.60
Deepest LLM values repeated, WISE	0.48	0.57
Local-edit comparison	**Instruction success**	**Unchanged LPIPS ↓**
Full model	79.2%	0.090
Reference depth reads disabled	71.7%	0.171
Trained without reference reads	76.7%	0.138

Table 19: **Denoising-window, knowledge and reference-read controls.** Top, the decoder uses fixed calibration-average attention during the first or last 10 Heun steps. Middle, the deepest LLM values replace intermediate ones, with and without the required facts in the prompt. Bottom, reference reads are disabled on 120 local edits or removed during training.

replacement and background editing, selected before viewing outputs. Independent masks mark unchanged regions, excluding global geometric or style edits. LPIPS-Alex v0.1 (Zhang et al., 2018) uses its default input range, and masks are eroded by 16 input pixels and area-resized to each feature grid. Two condition-blinded raters judge requested-change success. Success is 95/120 with full reads, 86/120 with reference reads disabled and 92/120 for a model trained without them.

G LIMITATIONS

LayerBridge re-encodes the evolving target with the frozen VLM at every sampling step, which keeps guidance current but makes sampling cost grow with the VLM (Section B). The compact 512^2 pixel decoder limits small text and fine textures, as the lower ImgEdit scores for extraction and hybrid edits show. Our experiments also use only the Qwen3.5 family and single-turn inputs.