V-ENGRAM: TRIGGER-INDEXED EXTERNAL MEMORY FOR MODULAR TEXT-TO-IMAGE PERSONALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained text-to-image models contain broad visual knowledge, yet they cannot reliably acquire or refine a specific visual identity from only a few references while preserving compositional control. Token-embedding methods are compact but often underfit identity, whereas adapter-based methods improve fidelity through persistent weight updates that can be costly to store and interfere when concepts are composed. We introduce V-Engram, a trigger-indexed external memory mechanism for Stable Diffusion 3.5. Each concept is assigned an explicit trigger that retrieves concept-specific memory, whose gated directions enter frozen text-encoder and MMDiT context states as relative residuals. Separating this memory from backbone adaptation enables prompt-selective and multi-concept access without merging model updates. Experiments show that V-Engram broadly matches DreamBooth-LoRA in overall subject fidelity while showing advantages in settings such as contextual subject preservation. Prompt-matched loading retrieves only matched entries, reducing most additional adaptation-state loading for a single-concept query. Qualitative results further demonstrate paired-trigger composition and same-class separation, while prompts without registered entries retain the frozen model's base behavior. Together, these results establish trigger-indexed memory as a modular interface for adding targeted visual evidence without rewriting the generator.

1 INTRODUCTION

Large text-to-image generators encode broad visual and compositional knowledge (Rombach et al., 2022; Esser et al., 2024), yet their coverage is neither complete nor uniformly precise. A model may miss a newly observed subject, confuse similar instances, or render a familiar identity only approximately. Personalization thus spans both acquiring new concepts and selectively refining imprecise prior knowledge. This raises a broader question: where should targeted visual evidence be stored, how should a prompt retrieve it, and how can it enter generation without overwriting useful priors?

Few-shot personalization is a natural testbed for this question (Gal et al., 2023; Ruiz et al., 2023; Wu et al., 2025; Chen et al., 2025). From a handful of references, a method must bind a concept to a textual handle, recover its fine-grained appearance, and follow new contexts, attributes, and interactions. The evidence should also be selectively addressable: inactive when its handle is absent and composable without identity omission or attribute leakage (Huang et al., 2025; Yao et al., 2025; Peng et al., 2026). Figure 1 illustrates this objective for one toy car across different environments and actions.

Existing methods expose a capacity–modularity trade-off. Textual Inversion stores a concept in a small token embedding but often lacks capacity for detailed identity (Gal et al., 2023; Wu et al., 2025). DreamBooth and LoRA increase capacity by adapting generator pathways, yielding concept-coupled state whose composition can introduce interference (Ruiz et al., 2023; Hu et al., 2022; Peng et al., 2026). Encoder-based systems instead encode references at inference time for instant general-subject or face conditioning (Ye et al., 2023; Zhang et al., 2024; Li et al., 2024; Wang et al., 2024). Together, these methods leave room for memory richer than a single token embedding but still explicitly addressed by text and inactive outside that address. This is particularly relevant to Stable

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.

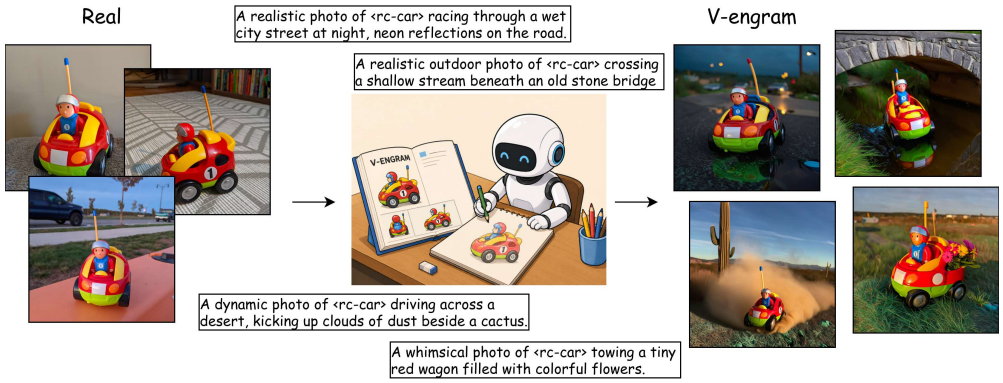


Figure 1: Compositional generation with V-Engram. A few reference images bind the trigger <rc-car> to a specific toy car; three are shown for illustration. The same memory composes with unseen environments, motion, lighting, and object interactions while preserving the characteristic driver, antenna, colors, and vehicle shape.

Diffusion 3.5, where text and image tokens interact throughout MMDiT blocks rather than through a fixed U-Net cross-attention interface (Esser et al., 2024; Peebles & Xie, 2023; Feng et al., 2026).

We propose V-Engram, a trigger-indexed external memory for Stable Diffusion 3.5. Each concept receives an explicit trigger such as <dog>, which registers tokenizer-specific n-gram entries. Exact matches activate the entries present in a prompt; their directions and tanh gates enter mounted text-encoder and MMDiT context states as residuals scaled to the local hidden-state norm. Concept evidence resides outside the backbone, and all pretrained components remain frozen.

Unmatched prompts receive no Engram residual; prompts containing registered triggers activate and sum all exactly matched entries. The interface can acquire a novel subject, separate similar instances, or refine an imprecise familiar concept. Storage grows with the registry, while inference activates only prompt-matched entries.

Experiments compare V-Engram with zero-shot Stable Diffusion 3.5, embedding-only learning, and joint DreamBooth-LoRA. V-Engram and LoRA show broadly comparable fidelity with complementary CLIP-I and DINOv2 outcomes. In held-out contexts, V-Engram preserves identity better, while LoRA favors prompt alignment. Qualitative studies show joint expression of separately addressed memories, inactive-route preservation, and same-class separation. Ablations support gated residual injection and distributed MMDiT access, while showing that denser mounting is not uniformly beneficial. Overall, V-Engram offers a different balance of fidelity, modularity, and compositional access rather than a universally stronger replacement for weight adaptation.

Our contributions are:

- We formulate few-shot acquisition and refinement as exact trigger-indexed memory retrieval over a frozen diffusion transformer.
- We introduce concept-specific Engram directions and gated relative residual injection for explicit, prompt-selective memory access in SD3.5.
- We characterize its trade-off against embedding-only learning and joint LoRA across fidelity, addressable loading, and compositional access.

2 Related Work

Conditional Memory. Parameter-efficient adaptation limits trainable state through adapters, prompts, or low-rank updates, but does not by itself specify concept-level activation (Houlsby et al., 2019; Li & Liang, 2021; Lester et al., 2021; Hu et al., 2022; Liu et al., 2022). External memory adds an addressing mechanism: Product-Key Memory performs structured content lookup, Memorizing Transformers use approximate kNN retrieval, and DeepSeek Engram deterministically maps matched token n-grams to learned entries (Lample et al., 2019; Wu et al., 2022; Cheng et al., 2026).

V-Engram shares deterministic lexical addressing, but changes both the memory content and target: few-shot visual references train concept-specific residual directions and gates for a frozen diffusion model, without semantic or approximate retrieval.

Optimization-Based Personalization. Learned concepts can reside in token space or model updates. Textual Inversion learns one embedding, CoRe regularizes its contextual interaction, and NeTI and ProSpect enrich frozen-backbone representations across layers or timesteps (Gal et al., 2023; Wu et al., 2025; Alaluf et al., 2023; Zhang et al., 2023). DreamBooth fine-tunes with prior preservation, while compact alternatives update attention, low-rank or rank-one factors, singular values, or sparse concept neurons (Ruiz et al., 2023; Kumari et al., 2023; Hu et al., 2022; Tewel et al., 2023; Han et al., 2023; Liu et al., 2023). CustomContrast and AttnDreamBooth further target subject discrimination or text alignment (Chen et al., 2025; Pang et al., 2024). V-Engram occupies a different point: the backbone remains frozen as in embedding learning, but its external, layer-wise memory is not confined to one input embedding.

Reference-Guided Methods. Image references can produce conditioning or compact adaptation state. BLIP-Diffusion, ELITE, IP-Adapter, and SSR-Encoder learn general subject features (Li et al., 2023; Wei et al., 2023; Ye et al., 2023; Zhang et al., 2024); PhotoMaker and InstantID specialize in human identity (Li et al., 2024; Wang et al., 2024); and HyperDreamBooth and AnyDoor predict compact weights or object features (Ruiz et al., 2024; Chen et al., 2024). These methods make a reference encoder or reference image part of the personalization interface. V-Engram instead distills references once into persistent, text-addressed memory, so generation is subsequently invoked by the trigger alone.

Multi-Concept Methods. Prior work jointly learns concepts, extracts several handles, or fuses adapters (Kumari et al., 2023; Avrahami et al., 2023; Gu et al., 2023). FastComposer and DynASyn target identity and action control, while condition merging and isolated sampling paths reduce multi-condition confusion and leakage (Xiao et al., 2025; Choi et al., 2025; Huang et al., 2025; Yao et al., 2025). TARA reduces interference in LoRA composition through rare-token masking and token–region alignment (Peng et al., 2026). V-Engram instead stores trigger-indexed direction-and-gate entries and injects them as exact-match hidden-state residuals, shifting the focus from spatially aligned LoRA composition to deterministic retrieval and selective activation of persistent concept memory.

Diffusion Transformers and MMDiT. Latent diffusion commonly uses a text-conditioned U-Net, whereas DiT replaces it with transformer blocks (Rombach et al., 2022; Peebles & Xie, 2023). Stable Diffusion 3.5 combines CLIP/OpenCLIP and T5 streams in MMDiT, where text and image tokens interact jointly (Radford et al., 2021; Ilharco et al., 2021; Raffel et al., 2020; Esser et al., 2024). Recent work uses DiT features for training-free subject injection, but MMDiT can still neglect or mix similar subjects (Feng et al., 2026; Wei et al., 2024). V-Engram therefore targets concept-specific text and context states inside SD3.5, beyond an initial embedding or U-Net cross-attention interface.

3 METHOD

3.1 OVERVIEW

Figure 2 shows the workflow. A few references train external entries registered from a trigger such as <dog>. At inference, exact tokenizer-specific span matches retrieve these entries; unmatched prompts follow the frozen path. Retrieved entries enter selected hidden-state sites as gated residuals.

Let G_θ be a pretrained Stable Diffusion 3.5 pipeline with frozen text encoders, VAE, and MMDiT transformer. A concept c is specified by a few reference images $\mathcal{D}_c = \{x_i^c\}_{i=1}^{N_c}$ and a trigger string s_c that appears in training prompts, e.g., "a photo of <dog>". The trainable concept memory ϕ_c consists of Engram directions and tanh gates attached to a set of mounted sites $\mathcal{Q}$. Each site $q = (g, \ell)$ denotes a tokenizer/conditioning stream g and a hidden-state location ℓ, with hidden width d_q; we write $g(q)$ for the stream associated with site q. The backbone parameters θ are never updated.

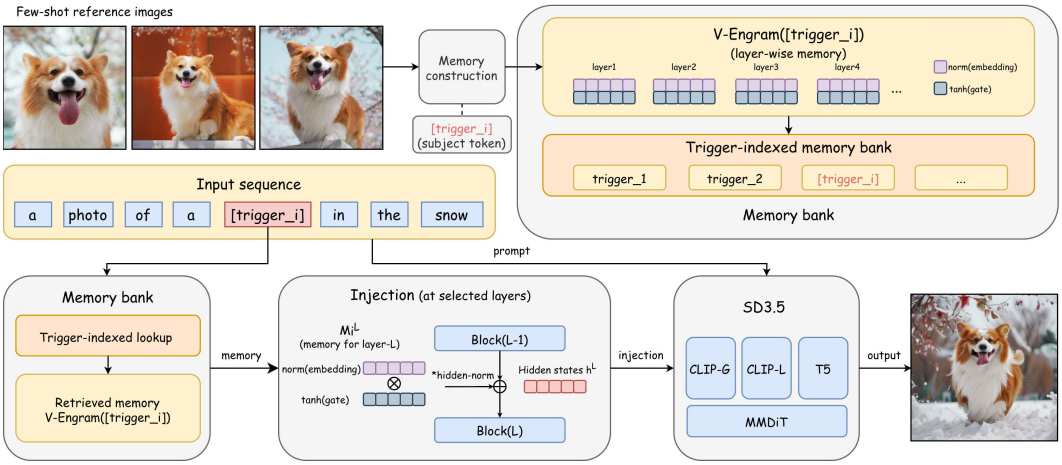


Figure 2: Overview of V-Engram. Few-shot references and an explicit trigger construct tokenizer-specific, layer-wise Engram entries. Exact token-id matches retrieve entries whose residuals are injected into selected frozen Stable Diffusion 3.5 layers.

3.2 Trigger-Indexed Memory Registry

Stable Diffusion 3.5 uses three tokenizer streams, CLIP-L, OpenCLIP-G, and T5-XXL; denote this set by $\mathcal{G}$. Since the same trigger can tokenize differently across streams, V-Engram builds a separate registry for each tokenizer. The registry for stream $g \in \mathcal{G}$ contains token sequences and the memory entries they address. For a prompt p, exact token-id scanning returns

$$\mathcal{A}_g(p) = \{(m, i) : \tau_g(p)_{i:i+|\kappa_m|} = \kappa_m,\ \kappa_m \in \mathcal{R}_g\}, \tag{1}$$

where κ_m is a registered span, m its Engram entry, and i its start position. The same surface trigger may tokenize differently across streams. The main configuration registers contiguous 2–4-grams and always retains the full trigger span; single-token triggers use exact spans. All exact matches, including overlaps, are active and summed below. This decomposition is intended to preserve local relations among constituent tokens in a multi-token trigger, a property not separately evaluated here. Lookup uses token ids rather than semantic or approximate retrieval, attention-based detection, or prompt rewriting. With no match, the conditioning path equals the frozen base model.

3.3 Gated Relative Residual Injection

For each registered entry m and mounted site $q = (g, \ell)$, V-Engram stores a trainable direction $e_{m,q} \in \mathbb{R}^{d_q}$ and a per-channel gate $a_{m,q} \in \mathbb{R}^{d_q}$. Routing is determined by the trigger lookup; the gate modulates the signed residual strength and is not a separate routing network. Since naive additive memories can over-perturb frozen hidden states, especially when mounted at several transformer blocks, V-Engram injects a gated relative residual. Given the hidden-state sequence H^q of length T_q at site q, the residual produced by entry m is

$$\begin{aligned} \Delta_{m,q}(H^q) &= \rho(H^q) \left(\frac{e_{m,q}}{\|e_{m,q}\|_2 + \varepsilon} \odot \tanh(a_{m,q}) \right), \\ \rho(H^q) &= \mathrm{sg}\left(\frac{1}{T_q} \sum\nolimits_{j=1}^{T_q} \|H_j^q\|_2 \right). \end{aligned} \tag{2}$$

The Engram direction supplies residual direction, the tanh gate provides bounded signed per-channel modulation, and the detached norm $\rho(H^q)$ calibrates the current activation scale. Our single-vector direct-embedding baseline disables this entire path and injects the learned embedding without a gate, direction normalization, or relative scaling:

$$\Delta_{m,q}^{\text{ungated}} = e_{m,q}. \tag{3}$$

For prompt p, let $\mathcal{C}_j^q(p)$ collect the matched entries whose trigger spans cover position j at site q. An entry m belongs to $\mathcal{C}_j^q(p)$ if its matched span in $\mathcal{A}_{g(q)}(p)$ covers token position j. Omitting the

repeated argument of $\Delta_{m,q}(H^q)$, the update is

$$\tilde{H}_j^q = H_j^q + \sum\nolimits_{m \in \mathcal{C}_j^q(p)} \Delta_{m,q}. \tag{4}$$

This form covers both the full trigger span and the case where a trigger such as <dog> activates several registered n-gram entries over the same token region. For CLIP-L and OpenCLIP-G, the injection is mounted before the penultimate transformer block, so downstream CLIP states used by Stable Diffusion 3.5 are Engram-conditioned. For T5-XXL, it is mounted before the final encoder block and affects the final T5 hidden states.

3.4 Memory Mounting in Stable Diffusion 3.5

V-Engram first attaches memory to the three text-encoder streams and also mounts memory inside selected MMDiT text/context blocks. We denote the selected MMDiT blocks by $\mathcal{L}_{\text{mmdit}}$. Unless otherwise specified, our default configuration uses seven evenly distributed blocks, $\mathcal{L}_{\text{mmdit}} = \{0, 6, 12, 19, 25, 31, 37\}$. Encoder-only mounting and different MMDiT mounting densities are evaluated in the ablation study. After SD3.5 concatenates CLIP and T5 contexts, matched CLIP spans map to the CLIP context region and matched T5 spans map to the T5 region. We maintain tokenizer-specific offsets when constructing this concatenated context sequence, so each trigger match is injected only into the corresponding CLIP or T5 token region. The pooled projection path is left unchanged. This gives a simple way to vary how deeply concept memory enters the multimodal denoising computation without modifying the pretrained transformer weights.

3.5 Training Objective and Storage

V-Engram stores external memory outside the pretrained model as small state dictionaries: three text-encoder Engram banks and their gate banks, plus selected MMDiT text/context banks. The trigger configuration records the trigger strings, tokenizer-specific registered targets, mounting sites, and gate setting. At inference time, the registry identifies exact prompt-matched entries, and only these entries are activated. Multiple concepts are composed by matching multiple triggers and summing all matched residuals; no LoRA adapters are merged and no backbone weights are modified.

For a concept c, let $K_{c,g} = |\mathcal{M}_{c,g}|$ be the number of entries registered in stream g. The trainable storage scales with registered entries and mounted sites:

$$|\phi_c|_{\text{params}} = 2 \sum\nolimits_{q \in \mathcal{Q}} K_{c,g(q)} d_q, \tag{5}$$

where $\mathcal{M}_{c,g}$ is the set of entries registered for concept c in stream g and the factor 2 corresponds to one direction vector and one gate vector. The ungated direct-embedding variant stores only the embedding vectors, giving $\sum_{q \in \mathcal{Q}} K_{c,g(q)} d_q$ parameters.

Training freezes all pretrained components and updates only Engram directions and gates. For a clean latent z_0, Gaussian noise ϵ, and flow time t, define $z_t = (1-t)z_0 + t\epsilon$ and the SD3.5 rectified-flow target $v = \epsilon - z_0$. The objective is

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{z_0,\epsilon,t,p}\left[\left\| F_\theta(z_t, t, p; \phi_{\mathcal{A}(p)}) - v \right\|_2^2\right], \tag{6}$$

where $\mathcal{A}(p) = \bigcup_{g \in \mathcal{G}} \mathcal{A}_g(p)$ denotes the active prompt-matched Engram entries across the three tokenizer streams. Prompts without active triggers use the unmodified frozen conditioning path.

4 Experiments

4.1 Experimental Setup

Benchmarks and references. The main benchmark contains 15 subjects and 80 Google DreamBooth references (Ruiz et al., 2023). Same-class disambiguation uses seven dogs and 37 references from the same dataset. For familiar-identity refinement, we select 15 Celebrity-1000 identities (tonyassi, 2026) and form, for each identity, a five-image training set and a five-image-disjoint test set with visibly varied pose, hairstyle, illumination, and setting.

Baselines and training. We compare frozen Stable Diffusion 3.5, TI-style embedding learning (Gal et al., 2023), joint DreamBooth-LoRA, and V-Engram. All learned methods use the same references and joint 15-subject training. TI-style, both LoRA ranks, and V-Engram use 25k, 8k, and 20k steps at learning rates 10^{-3}, 10^{-4}, and 5×10^{-4}, respectively. All use batch size 4, AdamW, a constant schedule, bfloat16, and seed 42. By default, V-Engram uses gated memories in the text encoders and seven MMDiT blocks $\{0, 6, 12, 19, 25, 31, 37\}$. Rank 4 (5.895M parameters) is the fixed LoRA configuration for controlled analyses; rank 8 (11.790M) is additionally evaluated in the main and contextual comparisons as a capacity check near the 10.477M full V-Engram registry. All qualitative figures containing LoRA outputs use rank 4.

Inference and metrics. To isolate subject fidelity, the main comparison uses "a photo of `<trigger>`", with the subject-specific address substituted for `<trigger>`. Seeds 42–46 yield five 768×768 images per subject using 20 steps, guidance 4.5, the default scheduler, no negative prompt, and bfloat16. CLIP-I and DINOv2 use CLIP-L/14 and DINOv2-base cosine similarity (Radford et al., 2021; Oquab et al., 2024). Best-reference takes each generation's maximum over its references; all-pairs averages the complete generation–reference matrix. Both are averaged within subject and then macro-averaged, respectively tolerating viewpoint variation and measuring full-reference consistency. CLIP-T encodes the exact contextual prompt, including its angle-bracketed trigger. For familiar identities, we compute cosine similarity between ArcFace-style face embeddings (Deng et al., 2019) extracted with InsightFace's `buffalo_l` model pack, comparing generations only against held-out references and reporting valid-face coverage separately. Further details are supplementary.

4.2 Main Comparison

Table 1 tests matching weight adaptation while improving on input-level learning.

Table 1: Main comparison over 15 subjects; five generations per subject, then macro-averaged.

	Avg. best-ref		Avg. all-pairs	
Method	DINOv2 ↑	CLIP-I ↑	DINOv2 ↑	CLIP-I ↑
Zero-shot SD3.5	0.4274	0.6765	0.3501	0.6361
TI-style	0.4919	0.7188	0.4071	0.6813
DB-LoRA ($r = 4$)	0.8084	0.8682	0.6911	0.8290
DB-LoRA ($r = 8$)	**0.8167**	0.8875	**0.6930**	0.8439
V-Engram	0.7896	**0.8902**	0.6809	**0.8463**

Increasing LoRA from rank 4 to 8 raises best-reference CLIP-I by 0.0192 and DINOv2 by 0.0082. Against the parameter-near rank-8 adapter, V-Engram is within 0.0028 on both CLIP-I aggregations, while rank-8 LoRA leads DINOv2 by 0.0271 best-reference and 0.0121 all-pairs.



Figure 3: Qualitative comparison. Columns: reference, zero-shot SD3.5, TI-style, rank-4 DreamBooth-LoRA, and V-Engram.

Taken together, the two methods achieve broadly comparable subject fidelity under the simple prompt: V-Engram essentially matches rank-8 LoRA in CLIP-I, whereas LoRA is higher in DINOv2, so neither dominates across visual representations. Figure 3 illustrates the rank-4 qualitative comparison for three representative subjects. The base and TI-style outputs frequently drift at the instance level, whereas rank-4 LoRA and V-Engram more faithfully recover distinguishing appearance, including the clock's twin-bell silhouette and yellow numeral motif, the boot's fringe structure, and the gray sloth plushie's shape and facial details.

Resource scaling. We measure addressability as the registry grows. Joint rank-4 LoRA remains fixed at 5.895M parameters; nested V-Engram registries with 1, 7, and 9 triggers contain 0.639M,

4.985M, and 6.221M, crossing LoRA at nine. In a controlled A40 profile (10 warmup and 100 measured steps), V-Engram uses 0.940–1.004 GiB less peak allocated training memory but takes 14.9–29.0% longer per step. For a single-trigger query to a 15-trigger checkpoint, matched loading activates 1.061M rather than 10.477M Engram parameters, avoiding 89.87% of full-registry adaptation-state loading. This enables selective adaptation-state management without loading unrelated memories. Per-setting measurements are supplementary.

4.3 Ablation Study

We isolate injection depth and the gated residual. The density study jointly trains five concepts (28 references); all variants use 10k steps, five seeds per concept, and fixed settings. Encoder-only retains three text-encoder sites; the others additionally use $N = 3$–9 uniformly distributed MMDiT blocks.

Table 2: Mounting-density ablation. N is the number of MMDiT blocks in addition to encoder sites.

Mounting setting	DINOv2 ↑	CLIP-I ↑
Encoder only	0.6132	0.7991
$N = 3$	0.7612	0.8499
$N = 4$	0.7814	0.8719
$N = 5$	0.7595	0.8820
$N = 6$	0.7780	0.8894
$N = 7$	**0.8068**	**0.9007**
$N = 8$	0.7345	0.8692
$N = 9$	0.8058	0.8798

All mounted variants beat encoder-only on both metrics, while density is non-monotonic. Seven blocks perform best on this validation subset and define our default, without implying an architecture-independent optimum. Compared with encoder-only injection, mounted variants more consistently preserve the subject's distinctive shape and appearance, with visible variation across densities.

A matched rerun compares TI-like direct single-vector injection with the complete V-Engram residual—normalized directions, hidden-state-relative scaling, and per-channel gates—under fixed data, schedule, and sites. The full design raises DINOv2 from 0.7065 to 0.8113 and CLIP-I from 0.8262 to 0.8892. Figure 4 shows the corresponding qualitative variation.

4.4 Additional Analysis

We next test held-out contexts, paired memories, unmatched routing, same-class instances, and refinement of imprecise prior knowledge.

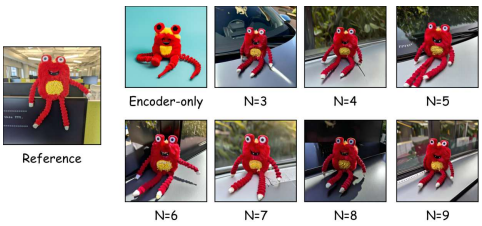


Figure 4: Qualitative mounting-density ablation; N denotes the number of selected MMDiT blocks.

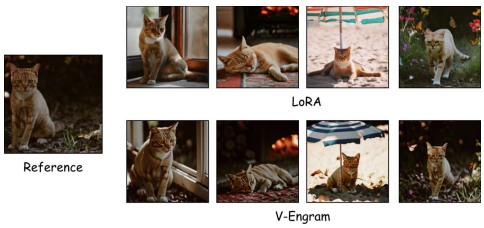


Figure 5: Additional contextual prompts for a personalized cat, comparing rank-4 DreamBooth-LoRA and V-Engram.

Contextual Composition. We apply five held-out, category-agnostic templates—close-up, room, snow, garden, and beside a red car—to all 15 subjects at seed 42. Shared wording enables controlled comparison across objects and animals without category-specific prompt engineering. Reusing the main checkpoints gives 75 images per method; Figure 1 separately shows broader composition for one toy car.

Table 3 shows that V-Engram better preserves personalized subjects in complex contexts, whereas rank-8 LoRA aligns text better. V-Engram leads rank 8 by 0.0612/0.0554 in best-reference/all-pairs CLIP-I (14/15 subjects), while rank 8 leads CLIP-T by 0.0324 (13/15). Raising LoRA from rank 4 to 8 improves CLIP-T by 0.0101 but reduces the two CLIP-I scores by 0.0158/0.0182; more capacity thus does not improve contextual fidelity. Figure 5, shown separately from the five-template protocol in Table 3, illustrates this trade-off. Its reference is at left; the upper and lower rows show rank-4 DreamBooth-LoRA and V-Engram, respectively. Exact prompts and additional examples are provided in the supplementary material.

Table 3: Contextual composition under five category-agnostic templates shared by all 15 subjects. Scores are macro-averaged over subjects.

	CLIP-T	CLIP-I	
Method	Score ↑	Best-ref ↑	All-pairs ↑
DB-LoRA ($r = 4$)	0.2558	0.8264	0.7888
DB-LoRA ($r = 8$)	**0.2658**	0.8105	0.7706
V-Engram	0.2334	**0.8717**	**0.8261**

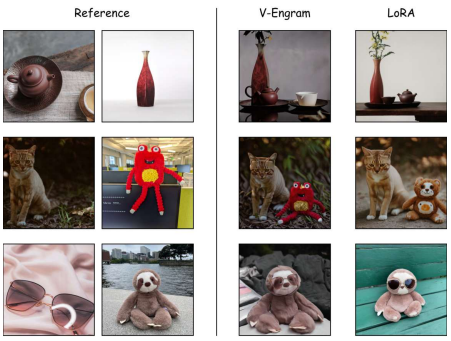


Figure 6: Paired-trigger composition for three concept pairs.

Multi-Concept Composition. Figure 6 shows three of 20 matched seed-42 paired-trigger prompts—teapot+vase, cat+toy, and sunglasses+plushie—against rank-4 LoRA. Its columns show two references, V-Engram, and rank-4 DreamBooth-LoRA. V-Engram more consistently depicts both concepts, while LoRA sometimes omits or blends one. We keep this comparison qualitative because full-image similarity cannot attribute retention to each subject; both methods still exhibit failures.

Non-Triggered Base-Behavior Preservation. Eight ordinary categories under seven templates yield 56 prompts without registered triggers. Exact-match logs show zero active entries in all three text streams. No residual is injected, so V-Engram exactly reduces to the frozen base under identical inference settings. Joint LoRA remains globally active and can shift outputs toward training instances. Supplementary examples visualize this difference.

Same-Class Trigger Disambiguation. Seven dogs (37 references) receive indexed triggers (<dog>, <dog2>, ...) or arbitrary names (<Oscar>, <Ryan>, ...). V-Engram trains for 10k steps and rank-4 LoRA for 4k. Seed-42 results cover a simple photo, running on grass, and sitting beside a red car; Figure 7 shows the red-car slice for four identities, with rows comparing V-Engram and rank-4 DreamBooth-LoRA under the two naming schemes. Because its exact-match registry accepts user-defined triggers, V-Engram is not tied to one naming convention; indexed and name-based triggers are the two forms tested here, and both better preserve dog-specific cues. LoRA tends toward similar dogs with indexed triggers and can invoke human priors with names. Without a calibrated dog-identity recognizer, this remains qualitative.



Figure 7: Same-class disambiguation for four dog identities.

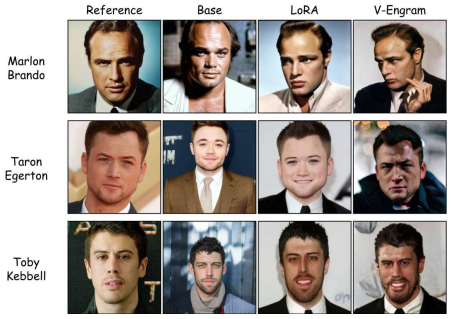


Figure 8: Familiar-identity refinement for three identities.

Refining Familiar Concepts. Finally, we test calibration of imprecise prior knowledge on 15 identities. Each condition generates five images per identity, compared only with five held-out references. The image-disjoint splits contain photographs from different times and capture conditions, rewarding recurring identity cues rather than training-image reconstruction. Figure 8 shows, from left to right, a held-out reference, zero-shot Stable Diffusion 3.5, rank-4 DreamBooth-LoRA, and V-Engram. V-Engram exceeds the tested rank-4 LoRA by 0.1466 best-reference and 0.1401 all-pairs similarity, winning 13/15 identities; both beat the base on all 15. All 75 generations per method are valid detections. The image-disjoint split supports identity-level refinement beyond single-image memorization, with three held-out visual examples shown in Figure 8.

Table 4: InsightFace similarity for 15 identities (generation-averaged, identity-macro-averaged).

	InsightFace		
Method	Best-ref ↑	All-pairs ↑	Valid
Zero-shot Stable Diffusion 3.5	0.2344	0.1824	75/75
DB-LoRA ($r = 4$)	0.4618	0.4014	75/75
V-Engram	**0.6084**	**0.5415**	75/75

5 Conclusion

We introduced V-Engram, a trigger-indexed external-memory interface that provides a frozen Stable Diffusion 3.5 backbone with prompt-addressed Engram entries. Under simple subject prompts, V-Engram is broadly competitive with LoRA, with comparable CLIP-I but lower DINOv2; in held-out contexts, it preserves identity better, while LoRA favors prompt alignment. Exact n-gram routing leaves unmatched prompts on the frozen base path and supports selective memory activation and loading; further analyses support familiar-identity refinement, paired-trigger composition, and same-class separation. Overall, these findings show a different balance among fidelity, modularity, and compositional access rather than universal superiority over weight adaptation. More broadly, exact token matching may provide an addressing mechanism for external memory in other token-conditioned generative or multimodal architectures, although cross-architecture validation remains future work.

AI Use Statement

Generative AI tools were used to assist with manuscript formatting and the conversion of the authors' existing submission package to the ICLR 2027 template. The authors reviewed the AI-assisted changes and take responsibility for the final content of this work, including all text, claims, and artifacts.

Reproducibility Statement

The appendix records the datasets and splits, trigger mappings, optimization settings, inference protocol, metric definitions, parameter accounting, and fixed qualitative selection rules used in the experiments. It also reports expanded per-subject results, ablations, failure modes, and the exact prompt banks used for contextual evaluation.

References

Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *ACM Transactions on Graphics*, 42(6):243:1–243:10, 2023. doi: 10.1145/3618322.

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 96:1–96:12, 2023. doi: 10.1145/3610548.3618154.

Nan Chen, Mengqi Huang, Zhuowei Chen, Yang Zheng, Lei Zhang, and Zhendong Mao. Customcontrast: A multilevel contrastive perspective for subject-driven text-to-image customization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2123–2131, 2025. doi: 10.1609/aaai.v39i2.32210.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6593–6602, 2024.

Xin Cheng, Rui Tian, Wangding Zeng, Damai Dai, Qinyu Chen, Bingxuan Wang, Zhenda Xie, Kezhao Huang, Xingkai Yu, Chengqi Deng, Shangyan Zhou, Chenggang Zhao, Zhewen Hao, Yukun Li, Han Zhang, Zhengyan Zhang, Yixu Wei, M. Y. Xu, Huishuai Zhang, Dongyan Zhao, and Wenfeng Liang. Conditional memory via scalable lookup: A new axis of sparsity for large language models. *arXiv preprint arXiv:2601.07372*, 2026.

Yongjin Choi, Chanhun Park, and Seung Jun Baek. DynASyn: Multi-subject personalization enabling dynamic action synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2564–2572, 2025. doi: 10.1609/aaai.v39i3.32259.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

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, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 12606–12633, 2024.

Haoran Feng, Zehuan Huang, Lin Li, and Lu Sheng. Personalize anything for free with diffusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3921–3929, 2026. doi: 10.1609/aaai.v40i5.37394.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. SVDiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7323–7334, 2023.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the International Conference on Machine Learning*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Qihan Huang, Siming Fu, Jinlong Liu, Hao Jiang, Yipeng Yu, and Jie Song. Resolving multi-condition confusion for finetuning-free personalized image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3707–3714, 2025. doi: 10.1609/aaai.v39i4.32386.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP. `https://github.com/mlfoundations/open_clip`, 2021.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Guillaume Lample, Alexandre Sablayrolles, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Large memory layers with product keys. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021.

Dongxu Li, Junnan Li, and Steven C. H. Hoi. BLIP-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2021.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked ID embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8640–8650, 2024.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, 2022.

Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. In *Proceedings of the International Conference on Machine Learning*, pp. 21548–21566, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Lianyu Pang, Jian Yin, Baoquan Zhao, Feize Wu, Fu Lee Wang, Qing Li, and Xudong Mao. Attndreambooth: Towards text-aligned personalized text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1259.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Yuqi Peng, Lingtao Zheng, Yufeng Yang, Yi Huang, Mingfu Yan, Jianzhuang Liu, and Shifeng Chen. TARA: Token-aware LoRA for composable personalization in diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8385–8393, 2026. doi: 10.1609/aaai.v40i10.37788.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6527–6536, 2024.

Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023. doi: 10.1145/3588432.3591506.

tonyassi. Celebrity-1000. `https://huggingface.co/datasets/tonyassi/celebrity-1000`, 2026. Hugging Face dataset, accessed July 2026.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. InstantID: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

Tianyi Wei, Dongdong Chen, Yifan Zhou, and Xingang Pan. Enhancing MMDiT-based text-to-image models for similar subject generation. *arXiv preprint arXiv:2411.18301*, 2024.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15943–15953, 2023.

Feize Wu, Yun Pang, Junyi Zhang, Lianyu Pang, Jian Yin, Baoquan Zhao, Qing Li, and Xudong Mao. CoRe: Context-regularized text embedding learning for text-to-image personalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8377–8385, 2025. doi: 10.1609/aaai.v39i8.32904.

Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations*, 2022.

Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 133:1175–1194, 2025. doi: 10.1007/s11263-024-02227-z.

Zebin Yao, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. Concept conductor: Orchestrating multiple personalized concepts in text-to-image synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9427–9435, 2025. doi: 10.1609/aaai.v39i9.33021.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics*, 42(6):244:1–244:14, 2023. doi: 10.1145/3618342.

Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, and Zhongliang Jing. SSR-Encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8069–8078, 2024.