SAEPER: Sparse AutoEncoders Projection for Reliable PERsonalization in Diffusion Models

Marek Skiba[1*], Jowita Drozdowicz[1*], Vladimir Zaigrajew[2], Przemysław Biecek[1,2], Piotr Sankowski[1]

[1] Uniwersytet Warszawski, [2] Politechnika Warszawska

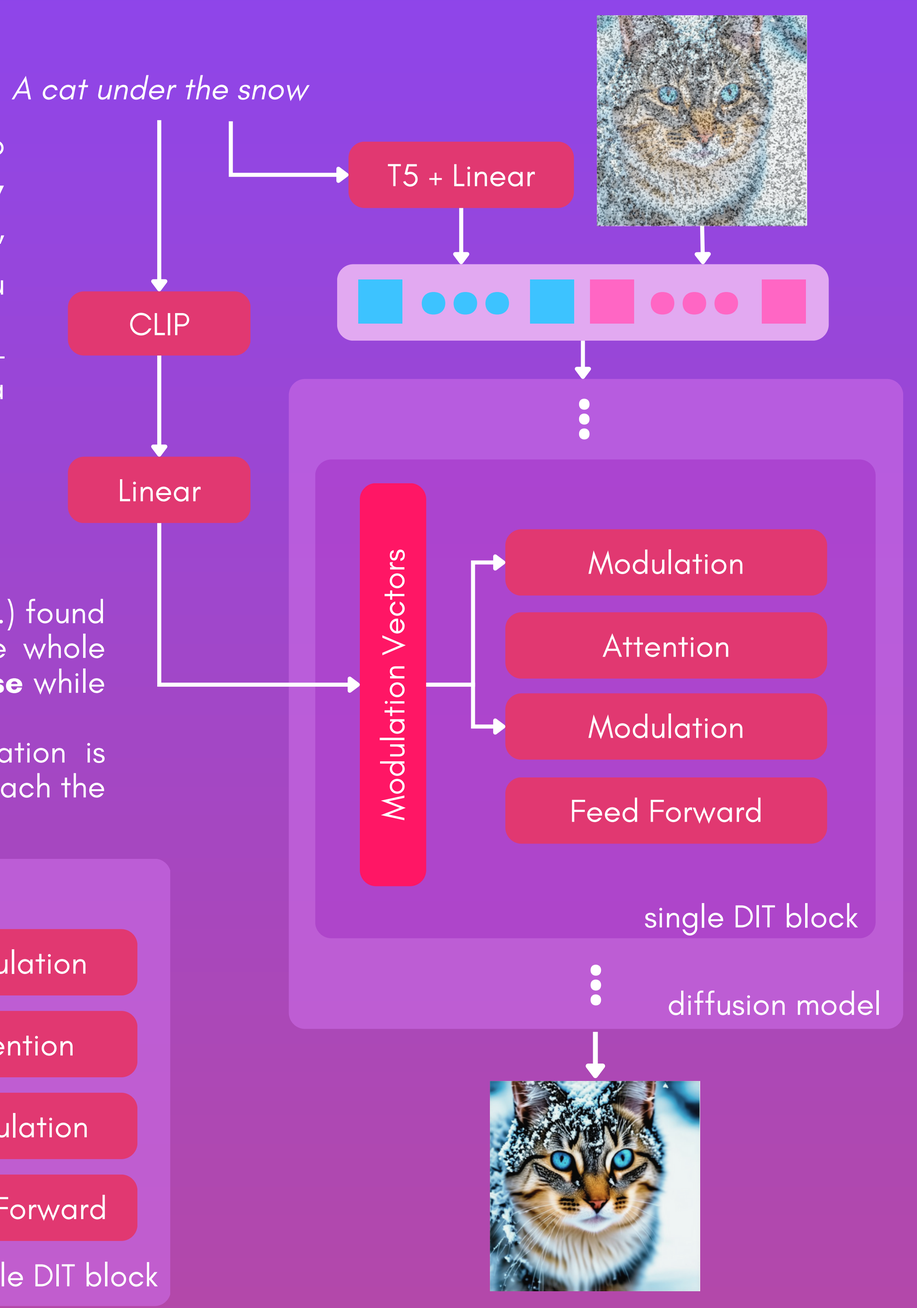
Personalization in Diffusion Models

One image, many concepts, endless possibilities

Personalization teaches diffusion models to **recreate a subject or attribute** from just one reference photo while keeping prompts fully open-ended. Instead of one monolithic image, we break it into **concepts: identity, material, style, lighting, pose**, and make each reusable in new scenes.

Our **lightweight controller** leaves the base model untouched: it modulates existing words (like "hat," "pose," "lighting") at inference. During training, it binds each visual concept to its corresponding word, letting you **recombine pieces of a reference photo through natural language**.

Generate a portrait with the same hat, move the pose to a beach, or keep the lighting but switch to watercolor—all without finetuning or new embeddings. It's **fast, compositional,** and **robust**, enabling faithful reuse from a **single image**.

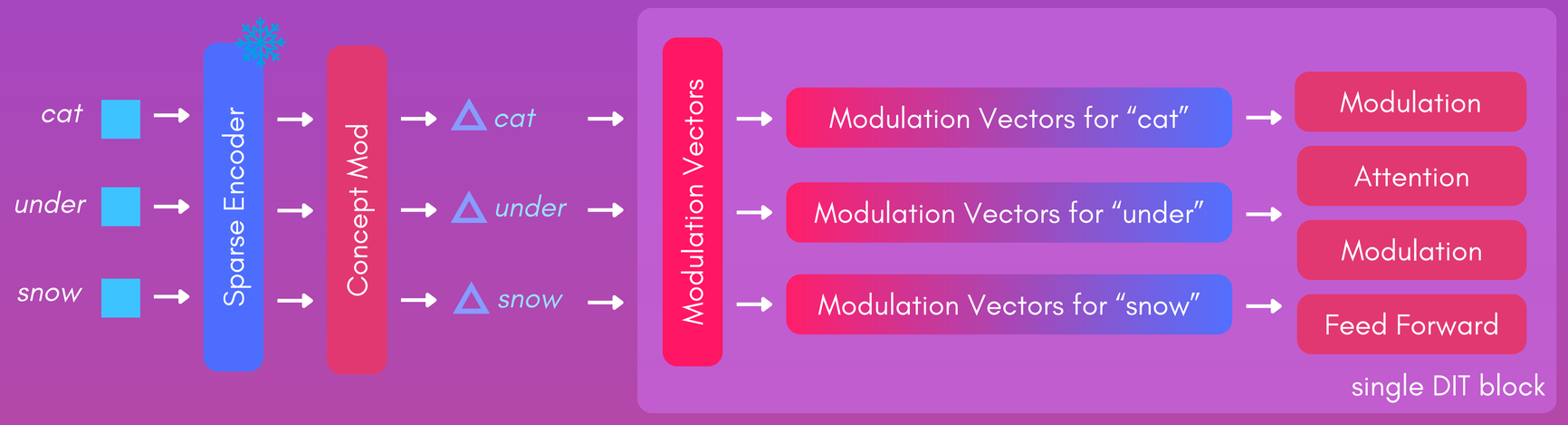


TokenVerse – Inspiration

One token, precise control, full freedom

Modern diffusion transformers guide images through **cross-attention** and **modulation**. **TokenVerse** (Garibi et al.) found that steering the modulation path per word token yields **precise**, **localized control**. Instead of altering the whole embedding, a **ConceptMod** module learns **per-token shifts** that adjust attributes like **identity, material, or pose** while keeping the rest untouched.

At inference, these **token-level steering directions** are plug-and-play—apply them only where personalization is desired. This enables **fine-grained edits**, **flexible composition**, and **zero base-model tuning**, a clean way to teach the model what to change and what to leave alone.



Reference image — **Generated images - SAEPER (ours)**

SAEPER – Our Method

Sparse codes, clean separation, stronger control

While TokenVerse offers control, **concept leakage** often remains: the "dog" still drags along its "hat." We address this by introducing **SAEPER**, combining TokenVerse with **Sparse Autoencoders (SAEs)** to **explicitly disentangle factors**.

We first encode tokens into **sparse, semantically independent features**, then let ConceptMod learn on this cleaner space. A **targeted pruning step** keeps only the most relevant neurons, ensuring compactness and focus.

This sparse representation **removes the need for isolation loss,** naturally enforcing **per-token separation** and minimizing leakage: **sharper control, cleaner composition,** and **higher visual fidelity** without sacrificing flexibility.

Experiments & Results

Sharper separation, faithful detail, robust composition

Our experiments show that **SAEPER** delivers a **significant boost in disentanglement** and **concept isolation**. Even **fine patterns** like stars on a backpack, jewelry details, or small fabric prints are correctly separated and can be **reused independently** in new prompts.

Unlike previous methods, SAEPER **preserves fine visual fidelity** while maintaining **clean per-token separation**, preventing attributes from leaking across concepts.

Reference image — **No SAE** — **With SAE** — **No SAE** — **With SAE**

Empirically, we observe **robust multi-image fusion** - concepts from several references can be combined into a single coherent scene, with each element retaining its distinctive look. SAEPer thus achieves **precise control** and **faithful reuse**, bridging creative flexibility with **consistent personalization**.

Infinite Composability

Independent concepts, infinite combinations composability

With clean separation of identity, clothing, and environment, SAEPER composes new scenes by **mixing elements across examples**: a hat, a pose, a pattern, a place, all staying faithful to their sources.

*: marek.skiba.poczta@gmail.com or j.drozdowicz@student.uw.edu.pl