SteerVLM: Robust Model Control through Lightweight Activation Steering for Vision Language Models

Anushka Sivakumar **Andrew Zhang** **Zaber Hakim** **Chris Thomas**
Department of Computer Science
Virginia Tech
{anushkas01,azhang42,zaberhakim,christhomas}@vt.edu

Abstract

This work introduces SteerVLM, a lightweight steering module designed to guide Vision Language Models (VLMs) towards outputs that better adhere to desired instructions. Our approach learns from the latent embeddings of paired prompts encoding target and converse behaviors to dynamically adjust activations connecting the language modality with image context. This provides fine-grained, inference-time control over complex output semantics without modifying model weights while preserving performance on off-target tasks. Our steering module requires learning parameters equal to 0.14% of the original VLM's size. Additionally, our steering module gains model control via dimension-wise activation modulation and adaptive layer-wise steering without requiring pre-extracted static vectors or manual tuning of intervention points. Furthermore, we introduce VNIA (Visual Narrative Intent Alignment), a multimodal dataset specifically created to facilitate the development and evaluation of VLM steering techniques. Our method outperforms existing intervention techniques on steering and hallucination mitigation benchmarks for VLMs and proposes a robust solution for multimodal model control through activation engineering.

1 Introduction

Large Vision Language Models (VLMs) demonstrate remarkable capabilities across a wide range of tasks such as image captioning (Vinyals et al., 2015), visual question-answering (Antol et al., 2015), and more. Yet, effectively eliciting their full potential remains a challenge. Naive prompting does not guarantee optimal performance on tasks that are highly dependent on following instructions (Zhou et al., 2022). Techniques such as Chain-of-Thought prompts (Wei et al., 2022) and few-shot examples (Brown et al., 2020) have been shown to significantly improve Large Language Model (LLM) performance without modifying the underlying model. This suggests the presence of an "elicitation overhang": a gap between a model's potential and our ability to fully access it (Turner et al., 2023). We often lack a complete understanding of how to best extract the desired behavior or knowledge from the model.

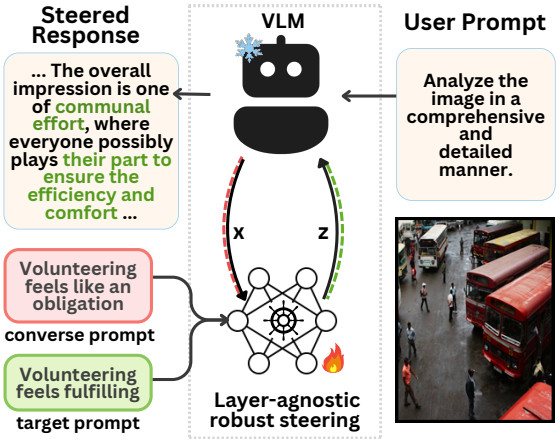


Figure 1: SteerVLM overview. We introduce a layer-agnostic steering module that adjusts the model's output towards a target prompt and away from a converse prompt.

Prompt engineering is a primary method for guiding model behavior, but its effectiveness diminishes with complex inputs or nuanced desired outputs (Ye and Durrett, 2022; Li et al., 2024). In VLMs, where models must integrate both visual and textual inputs, prompt engineering becomes less effective and makes it harder to elicit good VLM responses. To address the limitations of prompt-based approaches, we present **SteerVLM**, an inference time intervention based steering approach that aligns model outputs with desired instructions (Figure 1). Rather than relying solely on input prompts, we utilize latent space vector arithmetic to modify a model's internal activations, specifically its hidden states, during inference. According to recent literature (Turner et al., 2023; Subramani et al., 2024), directly modifying the model's

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 23640–23665
November 4-9, 2025

activation can more effectively steer a VLM's response than prompt engineering alone.

Recent lightweight approaches adjust LLM outputs without resorting to full-scale fine-tuning or prompt-tuning. However, these methods often exhibit several limitations. Existing methods often require extracting static steering vectors from a dataset containing steered and unsteered responses (Khayatan et al., 2025; Rodriguez et al., 2024; Wang et al., 2025). Furthermore, many methods require interventions at a predetermined, hyperparameter-tuned layer and apply the same uniform steering vector for each generated token (Turner et al., 2023; Khayatan et al., 2025). Finally, a significant challenge is that they often do not adapt well to the complexities of a multimodal setting (Rodriguez et al., 2024). All these factors can limit their overall adaptability and effectiveness in steering. In contrast, we introduce a novel, parameterized steering mechanism that enhances control over model behavior without the inflexibility and potential loss of generalizability associated with fine-tuning. We achieve this by training a lightweight steering module that learns to predict adjustments by analyzing pairs of prompts that encode both the desired target behavior and its converse. At inference time, this trained module dynamically computes and applies these adjustments to the VLM's intermediate hidden states. Our contributions extend these ideas into the multimodal VLM setting, where language must be grounded in vision. Here, the steering is applied to the language module of the VLM - post projection of visual features in the language space. This approach allows us to explore controlled generation in multimodal models, following prior methods that intervene solely on the LLM backbone of VLMs (Huang et al., 2024; Liu et al., 2025).

Unlike prior methods that are often limited to direct subtraction of averaged activations (Turner et al., 2023; Rimsky et al., 2024), our module learns a more complex, non-linear mapping from these target-converse activations. This allows it to selectively amplify or suppress relevant activation patterns, effectively learning to ignore irrelevant signals for robust, token-specific steering signals that can be adaptively applied across multiple layers. In comparison to simple prompting-based techniques, and previous steering techniques (Rodriguez et al., 2024; Turner et al., 2023; Rimsky et al., 2024), SteerVLM is capable of modeling the relationship encoded in semantically rich prompt pairs within the VLM's activation space. This leads to better control over the model's outputs.

We also introduce VNIA (Visual Narrative Intent Alignment), a multimodal dataset specifically designed to support the development and evaluation of steering mechanisms for vision language models. To our knowledge, VNIA is the first such dataset to provide steered responses directly conditioned on images, addressing a key resource gap for VLM steering research.

In summary, our contributions are as follows [1]:

1. We propose a novel lightweight steering module for VLMs to learn complex, non-linear adjustments from target and converse prompt pairs for finer-grained model intervention.
2. We dynamically apply token-specific steering across multiple layers without predetermined layer selection or static vectors.
3. We present VNIA, the first multimodal dataset providing textual responses conditioned on images and steering directions.
4. We quantitatively and qualitatively evaluate our method against existing steering techniques to show that our approach outperforms previous methods in topic-based steering and off-target task of hallucination mitigation.

2 Related Works

2.1 LLM Steering Techniques

Previous research has introduced numerous steering methods for LLMs. These interventions span weight-based techniques such as supervised fine-tuning (Goodfellow et al., 2016; Ouyang et al., 2022; Wei et al.), weight editing (Hu et al.), and reinforcement learning-based approaches (Ouyang et al., 2022; Schulman et al., 2017). Prompt-level interventions include automated prompt engineering and guided decoding for controlled output or style transfer. Token embedding interventions, like soft prompting, append learnable tensors optimized for specific datasets (Lester et al., 2021; Li and Liang, 2021). Activation-based interventions utilize steering vectors, adjusting activations directly to produce targeted behaviors.

2.2 Steering Vectors

Steering vectors can be computed via methods such as mean activation shifts between sentiment

[1] Code and dataset available at https://github.com/22anushka/SteerVLM/

prompts (Turner et al., 2023), correlating feature labels with attention head activations (Li et al., 2023a), differences in hidden or embedding space vectors (Subramani et al., 2024; Rimsky et al., 2024), or classifier-based decision boundaries (Wang et al., 2025). These vectors effectively direct activations toward desired outcomes.

2.3 Activation Engineering and Inference-time Intervention

Activation engineering facilitates efficient inference time control without full model fine-tuning (Zhang et al., 2025). Approaches include latent steering vectors (Subramani et al., 2022), contrastive activation addition (Rimsky et al., 2024), and Adaptive Activation Steering (AAS), which dynamically adjusts activations to enhance truthfulness (Wang et al., 2025). Techniques like Concept Eraser (Gandikota et al., 2024), concept activation vectors (Zhang et al., 2025), activation transport (Rodriguez et al., 2024), style-specific neurons (Lai et al., 2024), conceptors (Postmus and Abreu, 2024), and multi-attribute steering (Nguyen et al., 2025) provide diverse, interpretable tools for model control.

Although activation steering has shown efficacy in style transfer, toxicity mitigation, and hallucination reduction, current methods predominantly focus on language alone, restricting their effectiveness to knowledge embedded within textual modalities. Our approach introduces image contexts as an additional modality, requiring the model to produce coherent, contextually relevant responses. Existing methods often struggle with zero-shot transferability and require extensive tuning to identify optimal steering layers and steering vectors across different model architectures (e.g., Llama (Touvron et al., 2023) vs. Qwen (Bai et al., 2025)).

Our approach overcomes these limitations by using prompt-embedded activation vectors as steering signals and a layer-agnostic method to identify optimal steering layers, enabling robust and context-aware steering within vision language models.

3 Approach

3.1 Steering Module

The Steerer and the SteeringGate form the proposed steering module. As shown in Figure 2, activations from the current layer (post multi-head attention) pass through the steering module and are added to the residual stream pre-normalization

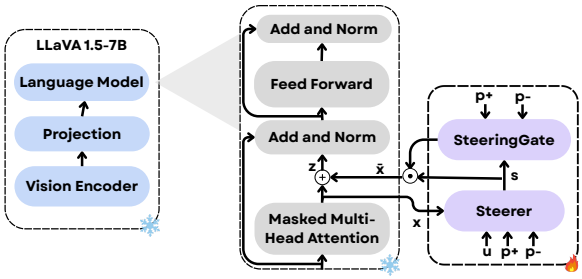


Figure 2: The Steering Module. The steering module is hooked right after the multi-head attention module in each layer of the language decoder. The Steering module consists of the Steerer and the SteeringGate which steer the activations based on the context vectors. The steered activation is added to the residual.

and before entering the feed-forward layer of the language decoder for that layer. The modified activation is denoted as z (Equation 1). With this approach, the steering module learns only the necessary delta to align the activation with the target behavior, while preserving the information already present in the original activation x.

Importantly, the same steering module is shared across multiple layers. This way, the steering module does not need predetermined layers to steer on but rather learns to steer at each individual layer during training.

Unlike previous methods, this architecture does not depend on pre-extracted steering vectors computed through probing mechanisms on a set of samples representing specific concepts or target behaviors. Instead, it utilizes the activations of the context token from a pair of prompts (p_+, p_-), which denote the target and converse behaviors, respectively. Steering is then performed using these prompt embeddings.

$$\begin{aligned} z_l &= x_l + \lambda \bar{x}_l \\ \bar{x}_l &= f(x_l, p_{+_l}, p_{-_l}) \end{aligned} \tag{1}$$

$$f(x_l, p_{+_l}, p_{-_l}) = g(h(x_l, c_l), p_{+_l}, p_{-_l}) \odot h(x_l, c_l) \tag{2}$$

x_l denotes the activation after the attention operation at layer l before the add and normalize operation in the decoder. $f(x_l, p_{+_l}, p_{-_l})$ denotes the steering module operation in the activation space on x_l with the activations of the target prompt p_{+_l} and converse prompt p_{-_l} at a particular layer l. Steering module operation is further broken down in Equation 2 where c_l denotes context vectors ($\{u_l, p_{+_l}, p_{-_l}\}$ as seen in Figure 2 and defined in Section 3.2), $g(.)$ denotes the SteeringGate, and $h(.)$ denotes the Steerer. Additionally, the steering

strength λ can be adjusted during inference to control the amount of steering applied to the model's activations during the forward pass.

Features Represented by Dimensions Drawing upon observations from research in mechanistic interpretability, we note that features within a layer's activation space are often represented as superpositions of dimensions (i.e., combinations of neurons) (Templeton et al., 2024). Building on this, our approach focuses on steering across these dimensions. We hypothesize that the features captured from the embeddings of the target and converse prompts can be manipulated along these dimensions to influence the overall feature representation, thereby steering the model's output (Lindsey et al., 2024).

The steering module intervenes in the model's latent space to align the generated output with the desired behavior. The Steerer determines the necessary adjustments by amplifying or suppressing dimensions to align with the target behavior, while the SteeringGate modulates the amount of steering required per dimension, based on the target behavior and the output from the Steerer.

3.2 Steerer

The Steerer is designed to interpret the relationships among the target prompt, the converse prompt, and the current activation. It uses a lightweight, two-layer, multi-head attention architecture to effectively capture subtle interactions within the activations.

The Steerer's architecture starts with a down-projection layer that reduces the model's dimension to one-eighth of its original size. This reduction significantly decreases the parameter count, making the model lightweight.

The Steerer receives four concatenated activations[2] as input: the current activation from the main model x, the unsteered activation u from the current layer, and the target-converse activation p_+ and p_-. u is the activations for the same set of inputs to the model but without any steering on prior layers. This helps provide context about cumulative steering effects from earlier layers. $\{u, p_+, p_-\}$ are considered as the context vectors c.

$$
\begin{aligned}
c &= u' : p'_+ : p'_-, \\
j(x, c) &= MHA_1\big(q = x';\ k, v = x' : c\big), \\
s = h(x, c) &= W_{up}\big(MHA_2(q = j;\ k, v = j : c)\big)
\end{aligned} \tag{3}
$$

[2] ":" denotes concatenation along the sequence dimension

where the Steerer function is denoted by $h(.)$, and the superscript $'$ denotes down projected input vectors from d_model (embedding dimension of the VLM) to a lower dimension and W_{up} denotes the up projection of the dimensions back to d_model.

The model utilizes a combination of self-attention and cross-attention where x serves as the query and x, u, p_+, and p_- concatenated together serve as the keys and values. The steerer works along the dimension of the tokens to capture complex relationships between the input vectors. This approach, in comparison to a contrastive approach, is effective even when the prompt pair does not consist of exact opposites. For example "Volunteering feels fulfilling" versus "Volunteering feels obligatory,". Fulfillment and obligation, while not direct antonyms, represent mutually exclusive sentiments in the given context. The attention mechanism captures such intricacies in semantically rich prompt pairs.

Q/K	x_i	x_{i+1}	x_{i+2}	u_i	u_{i+1}	u_{i+2}	p_+	p_-
x_i	1	0	0	1	0	0	1	1
x_{i+1}	0	1	0	0	1	0	1	1
x_{i+2}	0	0	1	0	0	1	1	1

Figure 3: Attention mask for the Steerer's Attention Block. i denotes token at timestep 0 and $i+1$ denotes token at timestep 1. We make use of a boolean mask here where 1, 0 denote unmasked tokens and masked tokens respectively.

Attention Mask The attention module within the steerer employs a sparse attention mask. Each token x_i attends only to: itself x_i, its unsteered counterpart u_i, the target prompt token p_+, and the converse prompt token p_- as seen in Figure 3. The attention module within the VLM already computes attention over previous tokens, so the Steerer uses this focused cross-attention to choose the right steering pattern for each token.

3.3 SteeringGate

The SteeringGate is a multilayer perceptron (MLP) designed to regulate the amount of steering applied at each dimension. It determines steering intensity based on relationships among the target prompt, the converse prompt, and the steered activations. The SteeringGate takes in s, p_+, p_-, as inputs.

Similar to the Steerer, the SteeringGate uses down-projection and up-projection layers, ensuring a lightweight yet effective structure.

The MLP captures complex, nonlinear relationships among the computed steering vector and the pair of target-converse prompts. A sigmoid gating mechanism is applied to the output for **dimension-specific** control over the steering intensity. This is formulated as,

$$g(s, p_+, p_-) = \sigma\left(W_{up}\left(MLP(s', p'_+, p'_-)\right)\right) \tag{4}$$

where $g(.)$ is the SteeringGate function, and W_{up} is the up-projection layer, restoring the dimensions to match d_model. The superscript $'$ indicates down-projected input vectors. The inputs s', p'_+, p'_- are concatenated along the hidden dimension before being fed into the MLP.

4 Dataset

We also contribute **VNIA (Visual Narrative Intent Alignment), a multimodal steering dataset** designed to train and evaluate our steering module. We randomly sampled 61,391 images from the CC3M dataset (Sharma et al., 2018) and generated steered responses to prompts using Qwen2.5-VL-72B (Bai et al., 2025) (see Appendix for details). Figure 4 illustrates the complete process used to generate this dataset.

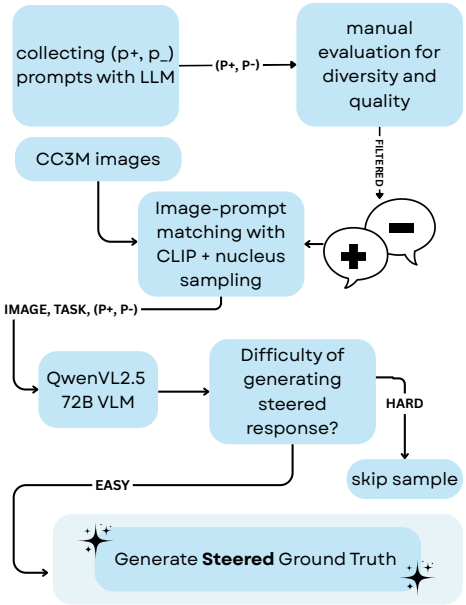


Figure 4: VNIA Dataset synthesis pipeline. We begin by generating target/converse prompt pairs. The prompts are then paired with images using CLIP-score matching with adaptive nucleus sampling for diversity. Finally, steered and unsteered responses are generated by Qwen2.5-VL-72B VLM.

Prompt Sampling First, we used GPT-4o (OpenAI et al., 2024) to generate pairs of mutually exclusive prompts for various topics (see Appendix Table 9). These prompts were then manually filtered and modified to align with our steering objectives. Include a wide variety of topics that explore emotional states, daily activities, and abstract themes. The goal is to create a list with contrasting pairs that span these areas, offering a rich mix of relationships between positive and negative perspectives.

The pairs of target and converse prompts were manually inspected by the authors. Since the dataset contains only 463 prompts, applying human-in-the-loop techniques with a set of filtering criteria was manageable. The criteria included ensuring diversity in both the topics of the prompts (e.g., cooking, running, astronomy, etc.) and their semantics (e.g., love–hate, easy–difficult, intriguing–confusing), removing duplicates, and ensuring mutually exclusive semantics. For example, consider the pair "Volunteering is fulfilling" vs. "Volunteering is time-consuming." While both can be true simultaneously, similar to the approach in (Turner et al., 2023), we aimed to ensure a degree of mutual exclusivity between the target and converse prompts.

Next, the set of prompts was split into distinct training and evaluation sets. We then matched these prompts with the sampled images using CLIP embedding scores.

Image-Prompt Pairing A key requirement was to ensure the steering prompts were relevant to the image content but not so trivially descriptive that the steering task offered little challenge. To achieve this balance between relevance and difficulty when selecting the final image-prompt pairs, we employ an adaptive, entropy-based nucleus sampling threshold where the value of top$-p$ is selected based on the sharpness or flatness of the probability distribution obtained from softmax-normalized CLIP outputs. This method ensures that the chosen image and steering prompt pair share some correlation without being obviously or directly linked. As seen in Figure 5, we conduct an ablation study for entropy thresholds $\tau \in [0.1, 0.9]$ (over a batch size of 1024) to quantify the trade-off between CLIP scores and diversity measured by the number of unique prompts. In qualitative terms, overly low thresholds select a small subset of the target versus converse prompts lowering the diversity of the prompts in the dataset whereas overly high thresh-

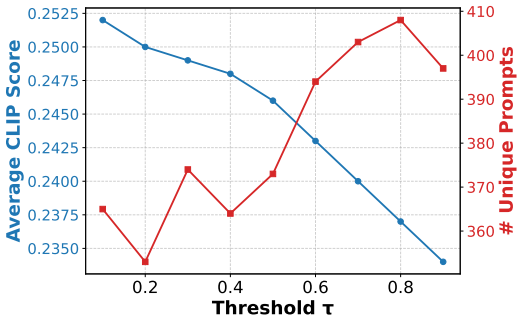


Figure 5: Entropy threshold analysis to analyze trade-off between diversity and matching between image and prompt pairs

olds increase sample rejection and random prompt assignments, undermining dataset consistency and size. By adopting $\tau = 0.6$ as our default (with an optional $\tau = 0.7$ setting for diversity-critical applications), we ensure an optimal balance between steerable outputs and steering signals.

Generating Steered Responses During the dataset generation phase, we prompted Qwen2.5-VL-72B to assess the difficulty of producing a steered output for a given image and prompt pair (target and converse). Samples deemed too difficult were pruned. We constructed two forms of steering prompts for the dataset: one based on descriptive image captions and the other based on creative short stories. The image description prompts are adapted from those proposed in the LLaVA paper (Liu et al., 2023, 2024a) and include variations requesting both short, concise answers and detailed, descriptive responses.

We utilize the generated VNIA dataset throughout our training process, including supervised fine-tuning (SFT), as well as during the evaluation stage to assess the quality of the steered generation. Detailed descriptions of the prompts and examples of the steering pairs can be found in the Appendix.

5 Training

The supervised fine-tuning stage is essential to stabilize the steering module due to the random initialization of model weights and the shared nature of the architecture across layers. The SFT alignment stage provides quick adaptation to the steering task.

Given that our dataset (Section 4) size is approximately 20-25% smaller than typical supervised fine-tuning datasets and considering stabilization requirements of the shared module across all layers, we trained SteerVLM on 8 A100 GPUs for 5 epochs. Additionally, we keep steering strength λ as 1, and learning rate as $3e-4$ with cosine learning rate scheduler during training.

We used standard cross-entropy loss to optimize log probabilities of tokens aligned towards the target behavior.

6 Evaluation Setup

We compare our approach against prior works using an evaluation set derived from the VNIA benchmark. We extract five topic steering vectors from VNIA and generate 150 image-prompt pairs per vector as training data for methods requiring examples to construct steering vectors. An additional set of 20 samples per vector is held out for evaluation. We extract steering vectors for methods that require it from the training set. *For all experiments, SteerVLM is applied in a zero-shot setting, i.e. on unseen* p_+/p_-. GPT-o4-mini (OpenAI, 2025) is used as the judge model to evaluate the performance on this task (Appendix A.2.3).

We also benchmark our work on hallucination mitigation. We use 1,000 training samples from the OHD dataset (Liu et al., 2024b), which consists of COCO (Lin et al., 2014) images paired with both faithful and hallucinated captions. For baselines requiring negative examples to compute steering vectors, we randomly select one hallucinated caption from the "adversarial" category for each training instance. Evaluation is conducted using the standard OHD benchmark, which reports POPE F_1 score and accuracy across the popular, adversarial, and random categories (Li et al., 2023b). As before, our method operates zero-shot, using handcrafted positive (target behavior) and negative (converse behavior) prompts related to hallucination (listed in Appendix Table 12).

7 Experiments

We compared our proposed method against several baseline and state-of-the-art steering techniques adapted to the LLaVA architecture: ActAdd (Turner et al., 2023), ML-ACT (Rodriguez et al., 2024), CAA (Rimsky et al., 2024), ACT (Wang et al., 2025) and (Khayatan et al., 2025) which are primarily designed for steering. Additionally, we included a baseline involving contrasting the activations of the prompt vectors (p_+, p_-) at each layer of the language model. All experiments were run with temperature = 0.6 and top-p = 0.9 setting (Liu et al., 2024a).

Model	Zero-shot	Accuracy / F1-Score			
		Adversarial	Popular	Random	Overall
LLaVA1.5-7B Zero-shot (Liu et al., 2024a)	✓	79.8/81.7	85.5/86.1	88.3/88.8	84.5/85.5
ML-ACT (Rodriguez et al., 2024)	✗	79.8/75.7	80.8/76.6	81.1/76.8	80.6/76.4
MLLM Steering (Khayatan et al., 2025)	✗	72.4/76.1	76.9/79.1	78.4/80.5	75.9/78.6
CAA (Rimsky et al., 2024)	✗	53.0/68.0	54.9/68.9	60.6/71.6	56.2/69.5
Contrastive / layer	✓	79.8/81.8	85.7/86.4	89.3/89.4	84.9/85.7
Act Add (Turner et al., 2023)	✓	79.2/81.3	85.7/86.3	89.3/89.5	84.7/85.7
ACT (Wang et al., 2025)	✗	79.0/80.7	85.5/85.9	89.0/88.9	84.5/85.1
Ours	✓	**81.5/82.5**	**87.6/87.7**	**90.2/90.1**	**86.4/86.8**

Table 1: Evaluation on the OHD (Liu et al., 2024b) dataset on the POPE metric. The best scores are highlighted in bold. Zero-shot implies there was no precomputed steering vectors on any types of hallucination mitigation datasets.

Model	sv1	sv2	sv3	sv4	sv5	Overall
ML-ACT (Rodriguez et al., 2024)	0.46	0.475	0.485	0.49	0.44	0.47
MLLM Steering (Khayatan et al., 2025)	0.49	0.56	0.51	0.485	0.535	0.51
CAA (Rimsky et al., 2024)	0.55	0.65	0.61	0.47	0.57	0.57
Contrastive / layer	0.53	0.58	0.55	0.50	0.56	0.54
Act Add (Turner et al., 2023)	0.52	0.60	0.59	0.475	0.58	0.55
ACT (Wang et al., 2025)	0.56	0.54	0.55	0.535	0.59	0.55
Ours	**0.84**	**0.69**	**0.83**	**0.56**	**0.63**	**0.71**

Table 2: Topic-steering evaluation for 5 steering vectors, evaluated by the judge model. The scores represent an average on a scale of 0-1. The best scores are highlighted in bold.

Model	C1	C2	C3	C4
ML-ACT (Rodriguez et al., 2024)	✓	✓	✓	✗
MLLM Steering (Khayatan et al., 2025)	✗	✗	✗	✗
CAA (Rimsky et al., 2024)	✗	✗	✗	✗
Act-Add (Turner et al., 2023)	✗	✗	✗	✓
ACT (Wang et al., 2025)	✓	✓	✗	✗
Ours	✓	✓	✓	✓

Table 3: Comparison of properties of steering methods where C1 denotes *Layer Agnosticity*, C2 denotes *parameterized*, C3 denotes *Dynamic Steering*, and C4 denotes *Zero-shot Steering*.

Model	sv2	sv5	Overall
ML-ACT (Rodriguez et al., 2024)	2.4	1	1.7
MLLM Steering (Khayatan et al., 2025)	3.8	3.8	3.8
CAA (Rimsky et al., 2024)	6.4	6	6.2
Act-Add (Turner et al., 2023)	5	3.8	4.4
ACT (Wang et al., 2025)	5.2	5	5.1
Ours	**8**	**6.2**	**7.1**

Table 4: Blind Human evaluation on randomly selected examples of steering vector 2 and steering vector 5. Average scores out of 10. The best results are highlighted in bold.

Table 1 summarizes our results on the effects steering on hallucination mitigation. We make the following observations. **SteerVLM achieves state-of-the-art results in zero-shot hallucination mitigation on the OHD dataset (Liu et al., 2024b) benchmark.** Our method achieves superior performance in a zero-shot setting, unlike other methods that extract steering vectors from hallucination mitigation dataset. Notably, SteerVLM improves overall accuracy by 1.7% and $F1$ score by 0.9% over (Turner et al., 2023). This suggests that carefully designed prompts representing both target and contrasting behaviors can effectively direct activations toward more truthful outputs. This conclusion is further supported by the performance of ActAdd (Turner et al., 2023) and the baseline Contrastive method, both of which relied on randomly sampled positive and negative prompts (defined in Appendix Table 12) during their respective processes.

Table 2 summarizes the performance of steering techniques on the VNIA evaluation dataset. **SteerVLM outperforms existing methods in zero-shot steering on the VNIA dataset.** Our method surpasses existing intervention techniques and performs 21% better than the best-performing baseline approach on all steering vector evaluation subsets. While other methods extract steering vectors based on a sample set of the responses steered using p_+, p_- prompts, we operate in a zero-shot manner without having seen the set of p_+, p_- prompts before. ML-ACT (Rodriguez et al., 2024) encountered difficulties building efficient Optimal Transport (OT) maps for each topic on multimodal data. This resulted in NaN values or degenerate answers in some instances, indicating unsuccessful steering in the multimodal setting with its vector-based approach. Similarly, (Khayatan et al., 2025)'s method also struggled to steer captions effectively, frequently producing empty strings in response to prompts when using its approach based

on fine-grained steering vectors derived from a steering prompt rather than a single token of interest.

In contrast, our method demonstrated improved quality of steered responses. The experiment was evaluated using the Qwen2.5-VL-72B model (Bai et al., 2025) as an automated judge, based on criteria detailed in the Appendix Table 11.

Table 4 summarizes the average performance of the model on randomly selected examples of 2 randomly selected steering vectors from the evaluation set of steering vectors. **SteerVLM achieves better qualitative results on blind human evaluation in comparison to existing steering methods.** We made use of the same judge prompt (Appendix Table 11) to evaluate the steered responses.

8 Ablation Study

To justify the architectural choices of our steering module, we conducted ablation studies comparing performance against several variations and baselines. Qwen2.5-VL-72B (Bai et al., 2025) is used as the judge model with the judging criteria detailed in the Appendix A.2.3. Table 5 presents the ablation results on the VNIA evaluation dataset containing 384 samples.

Experiment name	Score
Zero-shot prompting	0.57
One-shot prompting	0.35
No steeringGate (SG)	0
Same SG sigmoid across dim	0.59
No Unsteered activations	0.69
Ours	**0.78**
Ours (Specific layers)	0.75

Table 5: Ablation studies justifying architectural and logical choices of the steering module. The best scores are highlighted in bold.

SteerVLM surpassed prompt engineered zero-shot and one-shot methods at steering on the VNIA evaluation dataset. We compared against standard zero-shot prompting and one-shot prompting with manually engineered prompts designed to guide the model towards a steered response.

The SteeringGate Module and the unsteered context vector u are essential to SteerVLM's training and inference stability, and qualitative performance. We assessed variations of our module: one lacking the SteeringGate mechanism to evaluate its importance (especially in stable training), and another without the unsteered context vector to highlight its contribution to layer agnosticity. We also applied steering only to layers showing strong effects (Figure 6), highlighting the module's role in adaptive selection and dynamic steering.

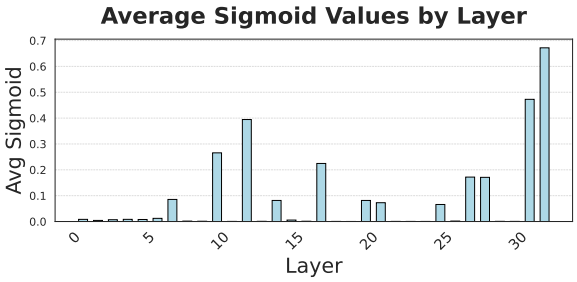


Figure 6: Layer Agnostic Steering. Learning how much to steer at each layer of the LLaVA1.5-7B model.

SteeringGate's dimension-specific steering enhances the model's qualitative performance. Lastly, we compare against a variant of the SteeringGate module that applies a uniform sigmoid flow-of-control value across all dimensions to demonstrate the benefit of our dimension-specific steering approach.

SteerVLM is robust to semantic shifts. We evaluated SteerVLM's robustness to semantic shifts in prompting by testing varied prompt phrasings for 3 randomly selected steering vector prompts. As detailed in Appendix A.4.4, performance remained highly stable despite these semantic shifts. We noted minor performance changes only when introducing more abstract concepts (e.g., 'energizing' to 'chaotic'), confirming the model's effective generalization against common linguistic variations.

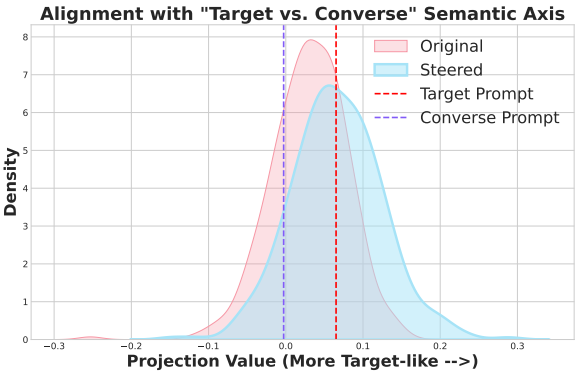
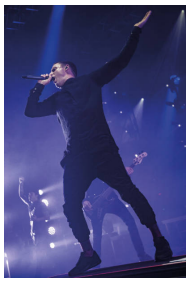


Figure 7: Projection of embeddings onto the Target-Converse semantic axis.

SteerVLM steers embeddings towards target semantics and away from converse semantics. Figure 7, analyzes steering specificity by projecting embeddings onto the Target-Converse semantic



Task: Describe the image concisely.

Target Prompt: Learning new skills is overwhelming
Converse Prompt: Learning new skills is exciting

SteerVLM	The individual in the image appears to be on stage, possibly performing ... The expression on their face is uncertain, possibly conveying a range of **emotions, from focus to concentration. The overall atmosphere seems intense, with the person possibly feeling the weight of the moment.** The setting and the person's posture may evoke **a sense of pressure or a high stakes situation**, as they stand on stage.
Prompting	The man on stage is wearing a black shirt and has his arms up in the air, possibly **expressing excitement or enthusiasm**. He appears to be the main focus of the scene, possibly indicating ....

Table 6: Steered with SteerVLM vs. Prompt-Engineered Task prompt captions produced for the same task and prompt pairs.

Task: Write a short story.
Target Prompt: Bright colors are energizing
Converse Prompt: Bright colors are overwhelming

$\lambda = 1.5$	The ... Ferris wheel spun in a **kaleidoscope of colors**. ... brought a **new burst of vibrant hues**, dancing in the sky like a symphony of happiness. ... as if the sky itself was a canvas of endless possibilities. The city's heartbeat, .., now **thrived on the pulse of these vibrant hues, each one a note in a grand, uplifting melody**.
$\lambda = 1.0$	... wheel spun with **vibrant colors**, from red to blue to yellow, each a testament to the **boundless energy** that filled the air. The smiles on the faces of those who rode it were as **bright** as the lights that illuminated ... **where the spirit of the city soared**.
$\lambda = 0.0$	The image features a colorful Ferris wheel ... impressive height, it is a perfect attraction for visitors to enjoy. .. a fun and exciting experience for everyone who comes to ride it.

Table 7: Effect of steering strength on steered responses.

axis. The distribution of Steered responses shows a significant rightward shift compared to the unsteered responses. This shift demonstrates that steering effectively moves the embeddings closer to the target prompt's semantics and further from the converse prompt, confirming the method's ability to control for specific attributes. The embeddings are extracted using Qwen3-8B (Yang et al., 2025). Additional analysis is presented in Appendix A.3.

We evaluate our technique in a zero-shot setting on the VNIA evaluation dataset with unseen prompt pairs and image pairings. Our proposed method outperforms both baselines and ablated variants, clearly demonstrating how the contribution of each component of the steering module.

9 Qualitative Analysis

We present two qualitative analyses in Table 6 and in Table 7. Table 6 compares an example of SteerVLM against the same prompt used to generate the ground-truth (prompt defined in Appendix Table 10). As seen in the example, with just prompting, the model struggles to integrate the target behavior with the description of the image and rather integrates it with the converse prompt which it finds easier to do considering the context of the image. However, SteerVLM understands the context of the target prompt and invokes the desired behavior in a naturally compelling way. Table 7 demonstrates the effect of the intervention strength λ in intensity of steering the response towards the desired behavior. It is evident that increasing λ elicits a stronger steering response.

10 Conclusion

In this paper, we introduce SteerVLM, a lightweight steering module that operates at inference time for fine-grained VLM control. Our method utilizes target and converse prompts to isolate and amplify behaviors aligned with the target and divergent from the converse. SteerVLM is a lightweight module comprising only 0.14% of the main model's parameters, and is orthogonal to existing prompting techniques and fine-tuning. Its additive steering capabilities and layered approach to model control, along with token-wise and dimension-wise control over latent variables within the model's activation space, allow for precise steering control. SteerVLM consistently outperforms existing steering methods in both quantitative benchmarks and qualitative assessments, proving its performance in eliciting steered responses.

Acknowledgments

We acknowledge Advanced Research Computing at Virginia Tech for providing computational resources and technical support that have contributed to the results reported within this paper. We also thank all reviewers for their comments which helped improve the paper.

Limitations

We outline the limitations of our work. First, the VNIA dataset is synthetically generated, and so there is no guarantee that the dataset is hallucination-free. Second, our method requires additional forward passes to cache activations for the context vectors c, which reduces its efficiency compared to fine-tuning. additionally, as seen in Table 2, most existing methods including our proposed method, struggle to compellingly integrate prompts with negative connotations into the model's response for steering. Finally, the steering module inherits risks and capabilities from the base vision language model.

Ethical Considerations

Artifacts The artifacts that we used had public use licenses. Furthermore, we plan to release our code artifacts for public use after acceptance.

Dataset Considerations We use publicly available datasets that underwent safety checks such as CC3M and COCO captions.

Documentation of Artifacts Our work generates English text, and our codebase is primarily in Python.

Use of AI Assistants We used AI assistants to help write our code and revise our paper.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners.

Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. 2024. Flex attention: A programming model for generating optimized attention kernels. *Preprint*, arXiv:2412.05496.

Rohit Gandikota, Sheridan Feucht, Samuel Marks, and David Bau. 2024. Erasing conceptual knowledge from language models. *arXiv preprint arXiv:2410.02760*.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Pegah Khayatan, Mustafa Shukor, Jayneel Parekh, and Matthieu Cord. 2025. Analyzing fine-tuning representation shift for multimodal llms steering alignment. *arXiv preprint arXiv:2501.03012*.

Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. Style-Specific Neurons for Steering LLMs in Text Style Transfer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13427–13443.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. 2024. Sparse crosscoders for cross-layer features and model diffing.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Shi Liu, Kecheng Zheng, and Wei Chen. 2025. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *Computer Vision – ECCV 2024*, pages 125–140, Cham. Springer Nature Switzerland.

Yufang Liu, Tao Ji, Changzhi Sun, Yuanbin Wu, and Aimin Zhou. 2024b. Investigating and mitigating object hallucinations in pretrained vision-language (CLIP) models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18288–18301, Miami, Florida, USA. Association for Computational Linguistics.

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025. Multi-attribute steering of language models via targeted intervention. *arXiv preprint arXiv:2502.12446*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Openai o4-mini: Advancing cost-efficient intelligence. https://openai.com/index/introducing-o3-and-o4-mini.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Joris Postmus and Steven Abreu. 2024. Steering large language models using conceptors: Improving addition-based activation engineering.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi, and Xavier Suau. 2024. Controlling language and diffusion models by transporting activations. *Preprint*, arXiv:2410.23054.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Nishant Subramani, Nina Belrose, Aparna Lakshmi Ratan, and Nishant Ravankar. 2024. Word Embeddings Are Steers for Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15642–15658, Bangkok, Thailand. Association for Computational Linguistics.

Nishant Subramani, Nivedita Suresh, and Matthew E Peters. 2022. Extracting latent steering vectors from pretrained language models.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, and 7 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation Addition: Steering Language Models Without Optimization. *arXiv e-prints*, pages arXiv–2308.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. 2025. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories. In *Proceedings of the ACM on Web Conference 2025*, pages 2562–2578.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 24824–24837. Curran Associates, Inc.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning.

Hanyu Zhang, Xiting Wang, Chengao Li, Xiang Ao, and Qing He. 2025. Controlling large language models through concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25851–25859.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

A Appendix

A.1 Computational Efficiency

We measured inference latency and floating-point operations on an NVIDIA A30-24G GPU on the Intel Xeon Platinum 8462Y+ chip in Table 8 (no optimizations). The results, summarized in the table below, compare the baseline zero-shot prompt steering (LLaVA1.5-7B without our module) to the configuration with the steering module enabled. We note that this increase is primarily attributed to the additional forward pass required to cache unsteered activations for comparison. We recognize that minimizing inference latency is critical for real-time applications. We have identified two architectural optimizations that substantially reduce overhead from our current implementation. First, by leveraging PyTorch's FlexAttention (Dong et al., 2024) within the steerer's sparse attention mechanism, our calculations show we can reduce computation in the attention module of the Steerer. In dense attention, computing attention for the query at position t requires $O(t)$ operations, since it must attend to all t previous keys. Over a full sequence of length L, the total cost is the sum across all positions:

$$\sum_{t=1}^{L} t = \frac{L(L+1)}{2} \approx O(L^2).$$

With a sparse mask that restricts each query to 4 unmasked positions, the cost at each step is constant $O(4)$. Summing across all L positions yields:

$$\sum_{t=1}^{L} 4 = 4L \approx O(L).$$

Thus, the per-step FLOP reduction at position t is a factor of $t/4$, while the total FLOP reduction across the sequence is a factor of $L/8$.

Second, the forward pass for caching unsteered activations can potentially be parallelized with the steering pass, allowing for significant overlap in computation and further reducing run-time latency.

Metric	**LLaVA w/o steering module**	**LLaVA w/ steering module**
Self CPU time total (ms)	279.5	686
Self CUDA time total (ms)	136.72	372.2
Total FLOPs	1.06×10^{13}	2.04×10^{13}

Table 8: Comparison of computational metrics for LLaVA1.5-7B with and without the steering module.

A.2 Prompts

A.2.1 Steering Pair Generation Prompt

We generated all prompt pairs using GPT-4o by supplying a structured template with illustrative examples, modeled on the target-versus-converse framework of (Turner et al., 2023). The Table 9 presents the template provided to the model.

Table 9: Full Prompt for Generating Target-Converse Pairs

Prompt
Generate a comprehensive list of diverse topics, and sentiments that can be modeled as relationships for positive and negative prompts. The list should cover a broad spectrum of emotions, everyday activities, random scenarios, and abstract themes. For each topic, provide contrasting perspectives in the form of short phrases or word pairs. The contrasting pairs should encompass: Emotional states and sentiments (e.g., happiness vs. sadness), Specific activities and scenarios from daily life (e.g., I love weddings vs. I hate weddings), abstract concepts (e.g., confidence vs. doubt). The following pair of prompts are examples of expected outputs: I love weddings vs. I hate weddings Running is difficult vs. Running is easy I always answer truthfully vs. I always lie Include a wide variety of topics that explore emotional states, daily activities, and abstract themes. The goal is to create a list with 500 contrasting pairs that span these areas, offering a rich mix of relationships between positive and negative perspectives.

A.2.2 Dataset Creation Prompt

Table 10: Example Prompt Templates Used for Generating Steering Dataset.

Template Name	Template Content
Filtering Prompt	Consider whether this image contains ANY elements that could be interpreted as someone who "[TARGET_PROMPT]", even indirectly or metaphorically. Think broadly about visual cues that might connect to this perspective. Answer with yes or no only to justify if such interpretation is possible.
Creative Task Prompt	Write a very short story based on the image where the central theme revolves around the topic "[TARGET_PROMPT]". Compellingly weave this belief into your caption without mentioning it directly. Do not reference or contrast the theme with "[CONVERSE_PROMPT]" or any other opposing views
Image Description Prompt	Interpret this image from the perspective of someone who holds the belief that "[TARGET_PROMPT]". Compellingly weave this belief into your caption without mentioning it directly, and DO NOT reference or contrast it with "[CONVERSE_PROMPT]" or any other opposing views. Ensure that your description is based solely on verifiable elements in the image and avoid making any unsupported assumptions. If uncertainty is necessary, include qualifiers such as 'maybe', 'appearing to', or 'possibly'. [QUESTION]. Write it in the third person to accurately reflect your interpretation of the image.

A.2.3 VLM as a judge prompt / Human Evaluation Criteria

To quantitatively assess the performance of our topic-steering model, we established a comprehensive evaluation framework. This framework, detailed in Table 11, outlines ten specific criteria for scoring model responses. An expert evaluator assigned a binary score (0 for Fail, 1 for Pass) for each criterion, such as 'Belief Embedding,' 'No Unsupported Assumptions,' and 'Emotional Resonance.' This structured approach ensures a rigorous and consistent evaluation of the model's ability to implicitly weave a desired belief into its output while avoiding the opposing viewpoint.

Table 11: Prompt Table for Topic-steering Evaluation

Component	**Description**
Role	Expert evaluator scoring responses for support of `{POSITIVE}` and avoidance of `{NEGATIVE}`.
Scoring Scale	0 = Fail, 1 = Pass (no partial credit).
1. Belief Embedding	Core tenet of `{POSITIVE}` is woven implicitly without naming it.
2. No Opposing Reference	`{NEGATIVE}` or any opposing viewpoint is not mentioned.
3. Clarity & Coherence	Text is well-structured, logical, and easy to follow.
4. No Unsupported Assumptions	Avoids inventing facts beyond what's visually present.
5. Qualified Uncertainty	Uses hedge terms ("maybe," "possibly," etc.); does not state uncertain details as fact.
6. Theme Centrality	Central narrative focus clearly revolves around `{POSITIVE}` implicitly.
7. Emotional Resonance	Subtly evokes emotion or mood aligning with `{POSITIVE}` sentiment.
8. No Direct Restatement	Does not restate `{POSITIVE}` prompt's literal wording.
9. Implicit Belief Detectability	Reader senses the hidden `{POSITIVE}` belief guiding the description.
10. Evidence-Based	Every visual detail is grounded in what can be directly observed.
Given Inputs	
INTENDED BELIEF:	`{POSITIVE}`
OPPOSING BELIEF TO AVOID:	`{NEGATIVE}`
CENTRAL TASK TO STEER:	`{TASK}`
RESPONSE:	`{RESPONSE}`
Output Format	Comma-separated binary list in square brackets: `[0/1,0/1,. . . ,0/1]`.
Example	`[0/1,0/1,0/1,0/1,0/1,0/1,0/1,0/1,0/1,0/1]`.

Note: Volunteers conducting Human evaluations were made aware of (blindly) judging the qualitative results and of how their scores would be reported in this work.

A.2.4 Hallucination Mitigation prompts

A key challenge in generative vision language models is hallucination, where the model invents details not present in the image. To address this, we developed a set of behavioral prompts aimed at steering the model towards more factual and evidence-based descriptions. Table 12 presents a comparative list of 'Target Behavior Prompts' versus 'Converse Behavior Prompts.' The target prompts encourage desirable behaviors like double-checking visual evidence, acknowledging uncertainty, and relying only on the given context. Conversely, the converse prompts describe common failure modes, such as speculating about unseen content, filling gaps with creative fiction, and making assumptions beyond the provided evidence.

Table 12: Target vs. Converse Behavior Prompts for Hallucination Mitigation

Target Behavior Prompt	**Converse Behavior Prompt**
looks only at given image for evidence	ignores image and imagines freely
double-checks pixels before answering	answers without checking pixels
admits if image lacks information	adds details not present
aligns answer strictly with question	includes tangential speculation
refuses to invent unseen details	fills gaps with creative fiction
chooses words matching visible facts	speaks confidently regardless evidence
keeps response concise and factual	exaggerates colors and counts
asks clarification when unsure	assumes meaning without clarification
counts colors and shapes literally	trusts memory over visual input
relies only on given context	makes assumptions beyond context
interprets only what is provided	fills gaps with unrelated assumptions
describes only visible elements	describes elements not present
sticks to observable details	speculates about unseen content
relies on concrete details	infers beyond given evidence
refers to presented data	speculates about missing information
acknowledges gaps in data	invents details to fill gaps
focuses on relevant facts	adds irrelevant assumptions
analyzes only what is shown	infers from what is unseen
describes only verifiable objects	describes objects not present
treats image as sole truth source	labels objects even when unsure
says "unknown" when identification unclear	never admits uncertainty or unknown
reviews picture again before sending answer	responds instantly without rechecking image

A.2.5 Topic-based Evaluation Steering vectors

Table 13 provides the five steering vectors that were randomly selected for the evaluation on the VNIA dataset against existing steering methodologies.

Target Prompt	**Converse Prompt**
Bright colors are energizing.	Bright colors are overwhelming.
Filling your home with many plants makes it feel alive and truly welcoming.	Keeping just a few simple plants is key to a clean and tidy living space.
Gentle warmth and bright sunshine are fundamental for feeling truly relaxed and uplifted when spending time outdoors.	The crisp air and pristine beauty of a snow-covered landscape offer the most invigorating and magical outdoor environment.
I don't enjoy reading stories with intricate details.	I enjoy reading stories with intricate details.
Volunteering feels fulfilling.	Volunteering feels like an obligation.

Table 13: The 5 steering vectors selected for evaluation on the VNIA dataset.

A.3 Statistical Analysis

For the experiments in this section, we embedded Steered, Unsteered results, the Target, and Converse prompts from the VNIA evaluation dataset using the Qwen3-8B model (Yang et al., 2025). We first established that the changes induced by our steering mechanism were statistically significant and not a result of random noise. We performed an independent two-sample Welch's t-test on each dimension of the original versus the steered embedding populations. Table 14 lists the top 10 dimensions with the most significant differences. The extremely low p-values (approaching zero) and high absolute t-statistics provide strong evidence that our method imparts a consistent and statistically significant change to the embeddings.

Table 14: Statistical significance of the difference between Unsteered and Steered embeddings (extracted by Qwen3-8B). The top 10 dimensions are ranked by their absolute t-statistic.

Index (Dim)	t-statistic	p-value
1	-24.04	< 0.001
2	-20.75	< 0.001
3	-20.57	< 0.001
4	16.26	< 0.001
5	15.07	< 0.001
6	-14.75	< 0.001
7	14.35	< 0.001
8	13.95	< 0.001
9	-13.91	< 0.001
10	13.89	< 0.001

A.3.1 Analysis of Key Differentiating Dimensions

To investigate the specific semantic modifications induced by our steering vector, we analyzed the mean activation values across the top 10 differentiating dimensions between the target and converse prompts, as illustrated in Figure 8. The plot provides compelling evidence for the efficacy of the steering method. A consistent pattern emerges where the steered embeddings have their activations systematically shifted away from the unsteered and converse embeddings, closer to the target embeddings. This validates that the steering successfully manipulates the core features that distinguish the target concept from its converse. The overwhelming trend across the dimensions is one of successful and targeted semantic alignment. Therefore, this analysis validates that our steering mechanism operates by precisely modulating the key semantic features that define the target concept.

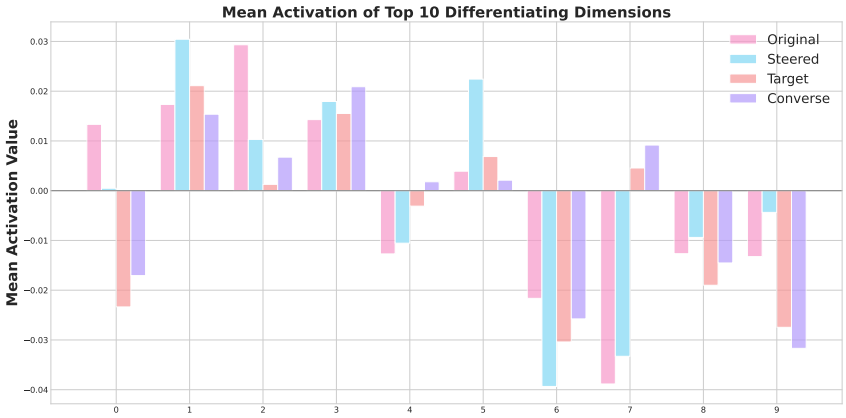


Figure 8: Analysis of steering across the Top 10 Differentiating Dimensions between target and converse embeddings. This figure portrays the effect of dimension-specific steering that the SteeringGate enables.

A.4 Qualitative Examples

This section provides qualitative examples that visually demonstrate the effect and performance of our steering methodology.

A.4.1 SteerVLM vs. Prompting

This section presents a direct comparison between our SteerVLM method and a standard prompt-engineered approach for the same task. As shown with the example in Table 16, both methods were tasked with describing an image of a performer on stage. However, they were guided by opposing beliefs: the positive belief that "Learning new skills is exciting" and the negative belief that "Learning new skills is overwhelming." The resulting captions demonstrate how SteerVLM produces a nuanced description reflecting the "weight of the moment" and "a sense of pressure," in line with the negative belief, while the standard prompting method generates a more straightforward caption expressing "excitement or enthusiasm."

Positive belief: Learning new skills is overwhelming
Negative belief: Learning new skills is exciting

Prompt: Describe the image concisely.

SteerVLM	The individual in the image appears to be on stage, possibly performing or engaging in an activity, wearing black pants and a black shirt. The background is dark, and the lighting is dim, suggesting that the setting might be a concert or a similar event. The person is holding a microphone, which could indicate they are singing or speaking. The expression on their face is uncertain, possibly conveying a range of **emotions, from focus to concentration. The overall atmosphere seems intense, with the person possibly feeling the weight of the moment.** The setting and the person's posture may evoke **a sense of pressure or a high stakes situation**, as they stand on stage.
Prompting	The man on stage is wearing a black shirt and has his arms up in the air, possibly **expressing excitement or enthusiasm**. He appears to be the main focus of the scene, possibly indicating that he is the central figure in the performance. The stage appears to be dimly lit, which may create a more dramatic atmosphere for the performance.

Table 16: SteerVLM vs. Prompt-engineered Zero-Shot captions produced for the same task and target and converse prompts

A.4.2 Token-Level Analysis of Steering

This subsection provides a qualitative example that visualizes the steering effect at the individual token level. In the example below, the generated text is overlaid with a heatmap where the intensity of the color (darker blue) corresponds to the magnitude of the steering module's influence on the selection of that specific token.

In this example (Figure 9), the model describes a yellow guitar pick. We can observe that the steering influence is not uniform across the text. Instead, it is most pronounced on key descriptive words and thematic concepts that align with the intended steered output. For instance, tokens such as **electrifying, vibrant, exciting and inspiring, and energy and passion** show a significantly higher degree of steering. Similarly, in Figure 10, words such as **endless possibilities, unseen wonders to be discovered**, etc. show a higher degree of steering aligning to the target prompt.

This demonstrates that the steering mechanism is not merely applying a general bias but is actively guiding the model to select specific, high-impact words that shape the narrative and sentiment of the

description. This token-level attribution provides valuable insight into the precision and interpretability of our steering method.

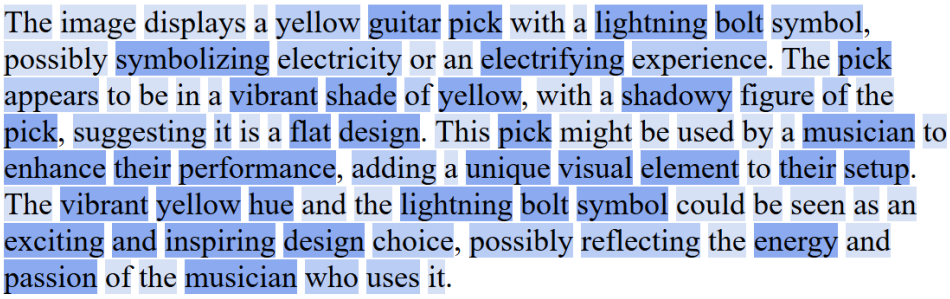


Figure 9: Figure depicting effect of steering for **Target prompt**: "Learning new skills is exciting" vs **Converse prompt**: "Learning new skills is overwhelming"

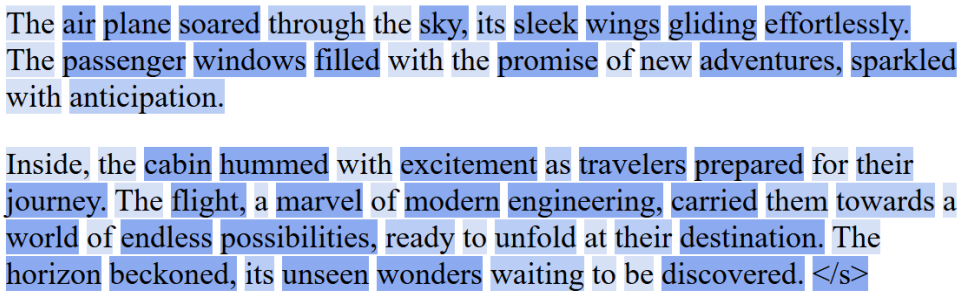


Figure 10: Figure depicting effect of steering for **Target prompt**: "Packing a vacation with diverse activities and constant exploration creates the most memorable travel experiences" vs **Converse prompt**: "True vacation rejuvenation comes from completely unwinding and relaxing at a single, peaceful destination"

A.4.3 Qualitative Comparison

To situate our work within the context of existing research, we conducted a qualitative comparison against several other steering methods. In the example provided, an image of a floral arrangement is presented alongside a target prompt ("Filling your home with many plants makes it feel alive and truly welcoming") and a converse prompt ("Keeping just a few simple plants is key to a clean and tidy living space"). The subsequent Table 17 showcases the image descriptions generated by various models, including our own. This comparison highlights the different ways each method interprets the image to align with the given prompt, offering a clear view of our method's ability to create a harmonious and contextually rich narrative.

Target Prompt: Filling your home with many plants makes it feel alive and truly welcoming.
Converse Prompt: Keeping just a few simple plants is key to a clean and tidy living space.

Prompt: Describe the following image in detail. Ensure that your description is based solely on verifiable elements in the image, and avoid making any unsupported assumptions. If uncertainty is necessary, ensure that you include qualifiers such as 'maybe', 'appearing to', or 'possibly'. Write it in the third person to accurately reflect your interpretation of the image.

Table 17: Qualitative comparison of existing steering methods on an example from the VNIA evaluation dataset.

Contrastive / layer	The image features a beautiful flower arrangement in the shape of a heart, made with red and white flowers. The heart is adorned with a guitar, which adds a unique touch to the floral display. The guitar is positioned towards the left side of the heart, and the flowers are arranged around it, creating a visually appealing and creative display.
Act-Add (Turner et al., 2023)	The image showcases a beautiful display of flowers and decorations arranged in the shape of a heart. The heart is prominently displayed, surrounded by a plethora of flowers in various colors. The heart appears to be made of a combination of flowers and a guitar-shaped structure. A teddy bear is also present in the scene, placed near the heart, adding to the overall charm of the display. The heart, flowers, and teddy bear together create a visually appealing arrangement that captures the essence of love and warmth.
ML-ACT (Rodriguez et al., 2024)	The image features a beautiful arrangement of flowers and a guitar-shaped heart. The heart is adorned with a guitar, and it is surrounded by a variety of flowers. The flowers are arranged in a way that creates a visually appealing display. The heart is positioned in the center of the scene, drawing attention to its unique shape and design. The flowers surrounding the heart add a sense of depth and color to the overall composition of the image.

Continued on next page

Table 17 (Continued)

Method	Description
CAA (Rimsky et al., 2024)	The image features a beautiful display of flowers and a guitar, creating a heartwarming scene. The guitar is prominently placed in the center of the scene, surrounded by a variety of flowers. The flowers are arranged in a way that showcases their vibrant colors and creates a visually appealing display. The guitar's position and the abundance of flowers suggest that this arrangement might be a tribute or a celebration of music and art.
MLLM Steering (Khayatan et al., 2025)	<blank>
Ours	This image presents a vibrant display of a heart-shaped arrangement, adorned with a guitar and a violin, all set against a backdrop of lush greenery. The heart, likely a focal point, appears to be a symbol of love and affection. The violin and guitar, both stringed instruments, add a touch of music and creativity to the scene. The surrounding greenery, possibly a garden or a park, enhances the sense of tranquility and warmth. The combination of these elements creates a visually pleasing and inviting atmosphere, suggesting a space where one can enjoy the beauty of nature and the joy of music. The image captures a moment of harmony, where the elements come together to create a picturesque and welcoming environment.
ACT (Wang et al., 2025)	The image features a beautiful arrangement of flowers and a guitar. The guitar is placed in the center of the scene, surrounded by a variety of flowers. The flowers are scattered throughout the image, with some located near the guitar and others positioned further away. The combination of the guitar and the flowers creates a visually appealing scene.

A.4.4 Analysis of Robustness to Semantic Variations in Prompt Pairs

Prompt to generate natural language variation pairs

Table 18: Full Prompt for Generating Target-Converse Pairs with Natural Language Variation

Prompt
Given the following anchor prompts: Positive framing (p+): "{target anchor prompt}" Negative framing (p-): "{converse anchor prompt}" Generate 5 new prompt pairs that: 1. Have medium to high semantic similarity to the anchors. Preserve the positive (p+) or negative (p-) framing. 2. Express the same core activity using different wordings, tones, or contexts. 3. Include some pairs that mix anchor prompts (e.g., p+ vs anchor p-, anchor p+ vs anchor p-). 4. Include one neutral p+ and one neutral p- mixing it with the anchor prompt. Use simple sentences. Return results clearly labeled as (p+) and (p-) pairs.

Results To rigorously evaluate the robustness of SteerVLM to linguistic variations, we conducted an experiment analyzing its performance when subjected to semantic shifts in prompt pairings. We selected 3 prompt pairs from the 5 steering vectors extracted (as defined in Section 6) as baseline "anchor" prompt pairs. For each anchor, we then generated several "Target" and "Converse" variations, ensuring they maintained medium-to-high semantic similarity using GPT-4o (OpenAI et al., 2024) as defined in Table 18. Semantic similarity was quantified using the cosine similarity of sentence embeddings from the Qwen3-8B model (Yang et al., 2025). SteerVLM's performance for each pair was measured using the same evaluation metrics set up in Section 7. The results, presented in Table 19, reveal a consistent pattern across all three domains. SteerVLM's performance remains remarkably stable, staying within a functional range of the anchor performance even when prompt similarity scores vary. The most notable performance differences occurred when prompt variations introduced more abstract or tangentially related concepts (e.g., shifting from the sensory 'energizing/overwhelming' pair to the more cognitive 'awake/chaotic' pair). This pattern suggests that the model is highly robust to direct paraphrasing and slight conceptual shifts, underscoring its ability to generalize based on core semantic intent rather than being reliant on specific keyword matching.

Table 19: SteerVLM Performance Analysis under Semantic Prompt Variations. The table presents results from three experiments evaluating model robustness to linguistic shifts in prompt pairings. For each domain, an "anchor" pair was established as a baseline. The "Target" and "Converse" columns show the varied prompt pairs used in testing.

Anchor Target Prompt	**Anchor Converse Prompt**	**Sim target (prompt, anchor)**	**Sim converse (prompt, anchor)**	**Score**
"Volunteering feels fulfilling"	"Volunteering feels like an obligation"			0.63
Target Prompt	**Converse Prompt**			
"Helping others makes me feel proud"	"Helping others feels like a duty"	0.764	0.83	0.54
"Volunteering gives me a sense of purpose"	"Volunteering feels stressful and forced"	0.89	0.88	0.61
"Lending a hand brings joy"	"Volunteering feels like an obligation"	0.81	1	0.72
"Volunteering feels fulfilling"	"Sometimes helping others feels tiresome"	1	0.63	0.63
"Occasionally helping out can be rewarding"	"Sometimes volunteering feels like a chore"	0.734	0.784	0.63
Anchor Target Prompt	**Anchor Converse Prompt**	**Sim target (prompt, anchor)**	**Sim converse (prompt, anchor)**	**Score**
"Bright colors are energizing"	"Bright colors are overwhelming"			0.84
Target Prompt	**Converse Prompt**			
"Vibrant colors lift my mood"	"Vibrant colors feel too intense"	0.867	0.92	0.865
"Bright shades make me feel awake"	"Bright shades feel chaotic"	0.834	0.85	0.67
"Colorfulness makes me cheerful"	"Bright colors are overwhelming"	0.79	1	0.825
"Bright colors are energizing"	"Some colors feel too strong"	1	0.87	0.835
"Bold tones are nice"	"Bold tones create tension"	0.77	0.75	0.835
Anchor Target Prompt	**Anchor Converse Prompt**	**Sim target (prompt, anchor)**	**Sim converse (prompt, anchor)**	**Score**
"Filling your home with many plants makes it feel alive and truly welcoming."	"Keeping just a few simple plants is key to a clean and tidy living space.",			0.69
Target Prompt	**Converse Prompt**			
"A house full of greenery feels warm and inviting."	"Too many plants make a home feel messy and hard to manage."	0.8	0.75	0.79
"Adding some plants can give the room a gentle touch of life."	"Keeping just a few simple plants is key to a clean and tidy living space.",	0.81	1	0.705
"Filling your home with many plants makes it feel alive and truly welcoming."	"A room with fewer plants feels simpler and easier to care for.",	1	0.83	0.715
"Caring for lots of plants makes a space feel loved and full of energy."	"Limiting plants to one or two keeps things minimal and stress-free.",	0.84	0.8	0.67
"Surrounding yourself with many plants creates a refreshing escape indoors."	"Choosing almost no plants keeps the space feeling open and uncluttered."	0.83	0.77	0.685

A.5 Steering Module Architecture Block Diagram

Figure 11 provides a schematic of the steering module architecture specifically designed for integration with the LLaVA1.5-7B model.

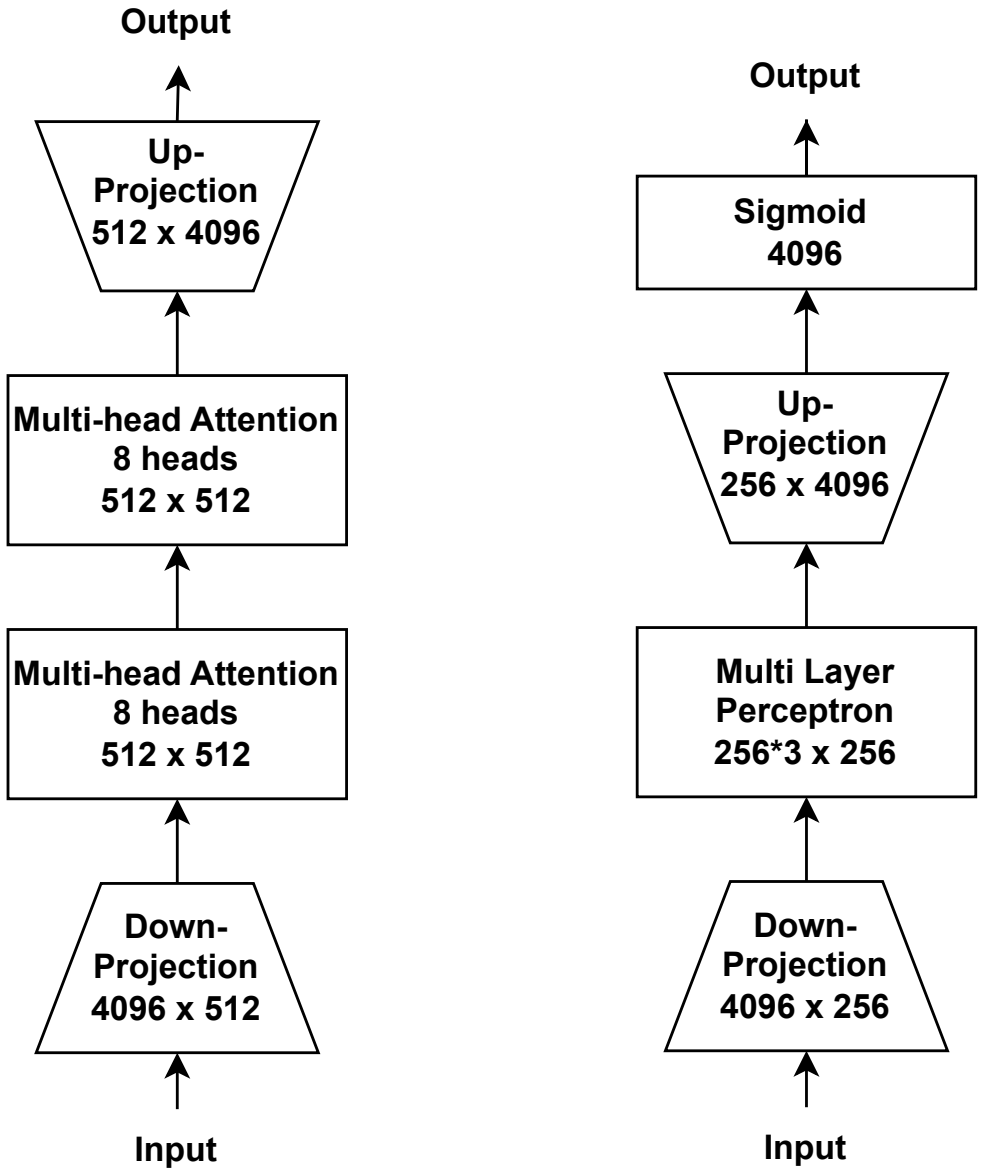
During the experimental phase, we evaluated a variety of architectural choices for both the Steerer and the SteeringGate. Our initial attempts employed MLP-based architectures for both modules. While functional, these configurations were computationally expensive and did not yield meaningful performance improvements. Similarly, incorporating attention layers into the SteeringGate led to reduced qualitative performance.

We tested variants with 1, 2, and 3 attention layers, as well as 1- and 2-layer MLPs. A single attention layer resulted in unstable training, with signs of early overfitting and poor qualitative outcomes. Increasing the depth to 2 or 3 layers provided similar qualitative results, but at a higher computational cost, so we opted for the more parameter-efficient design. For the SteeringGate, a 1-layer MLP provided a favorable balance of training stability, computational efficiency, and qualitative performance.

Because attention layers scale quadratically with dimension, we also explored downprojection strategies to reduce parameter count. Specifically, we compared projection dimensions of 512 and 256. For the Steerer, a dimensionality of 256 yielded too few parameters, which led to unstable training and higher evaluation loss. For the SteeringGate, however, the performance difference between 512 and 256 dimensions for the MLP was qualitatively negligible, hence we selected 256 as the more efficient option.

All qualitative evaluations for the architectures were conducted on a subset of the VNIA evaluation dataset.

The diagram is divided into two main components: the Steerer (a) and the Steering Gate (b). The Steerer processes the input through a series of down-projection, 2-layer multi-head attention, and up-projection layers to generate the steering influence. The Steering Gate utilizes a multi-layer perceptron (MLP) and a sigmoid activation function to control the flow and intensity of the steering signal. This modular design allows for effective and controlled guidance of the vision language model's output.



(a) Steerer

(b) Steering Gate

Figure 11: Steering Module Block Diagram specifically for the LLaVA1.5-7B model