Wiener Representation Filtering for VLM Hallucination Suppression

Ameen Ali*, **Tamim Zoabi***, **Lidor Brami & Lior Wolf**
Tel Aviv University
{ameenali,tamimzoabi,lidorbrami}@mail.tau.ac.il, wolf@cs.tau.ac.il

Abstract

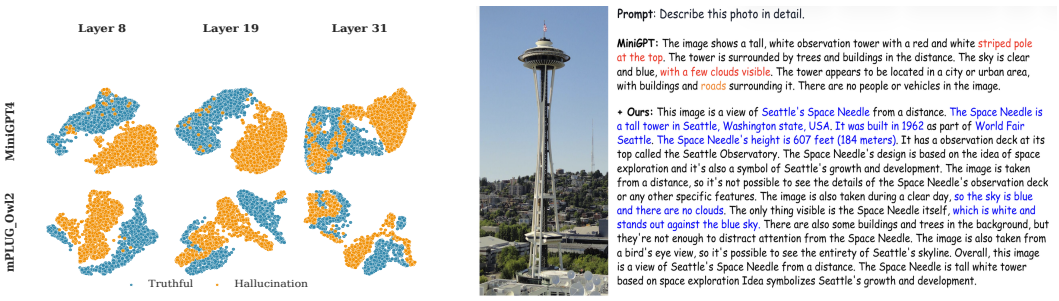
Vision-language models (VLMs) excel at open-ended captioning and visual QA but often describe objects, attributes, or relations absent from the image, a phenomenon known as object hallucination. We propose a training-free, post-hoc representation editing technique that operates in the representation space of the language backbone. The method performs a lightweight, one-time offline calibration on a modest paired dataset to estimate the required covariance structures, using only forward passes and empirical second-order statistics with no gradient updates or fine-tuning, after which the correction is absorbed directly into the model's existing weights. By modeling hidden states as a superposition of truthful and hallucination-associated components, we derive a Wiener-type estimator whose optimal gains are given in closed form from the covariances of paired truthful and hallucinated representations. An eigendecomposition yields mode-wise attenuation that respects a stability criterion, i.e., the filter responds continuously to estimation noise. The correction is applied once to the feed-forward output projections of selected deeper layers, at inference time, the model runs unchanged and at the same speed. Experiments on LLaVA-1.5, MiniGPT-4, Gemma3, and mPLUG-Owl2 demonstrate consistent reductions in object hallucination on CHAIR, POPE, and MME while maintaining caption fluency and overall response quality. We further demonstrate the generality of our approach on the TempCompass video understanding benchmark and on discrete diffusion language models for grounded dialogue, showing that representation filtering reduces hallucinations even in temporal video reasoning and multi-step, sequence-wide denoising settings.

1 Introduction

Vision-language models (VLMs) have achieved remarkable success in bridging visual perception and natural language processing (Liu et al., 2023; Ye et al., 2023b;a; Chen et al., 2023; Zhu et al., 2023), powering diverse applications such as image captioning, visual question answering, and multimodal reasoning. By integrating pre-trained vision encoders with large language models (LLMs), these systems generate fluent, context-aware descriptions of visual scenes. Despite their prowess, VLMs remain plagued by *object hallucinations* (OH): the fabrication of objects, attributes, or spatial relations absent from the input image, often stemming from linguistic priors or distributional biases in training data (Huang et al., 2024b; Rohrbach et al., 2018; Bai et al., 2024; Zhong et al., 2024; Huang et al., 2024b).

Existing mitigation strategies fall into two categories: decoding-time, such as constrained beam search (Freitag & Al-Onaizan, 2017a) or logit debiasing (Chuang et al., 2024; Leng et al., 2024), which modulate generation to promote grounded outputs, and post-hoc refinements such as Woodpecker (Yin et al., 2024), LURE (Zhou et al., 2024)), and the reference-guided editing method HALC (Chen et al., 2024). While these yield gains in controlled scenarios, they introduce substantial overhead, often 5–10x inference slowdowns, demand task-specific

*These authors contibuted equally to this work



(a) (b)

Figure 1: (a) UMAP visualization of MiniGPT4 and mPLUG-Owl2 hidden activations for truthful (blue) and hallucinatory (amber) samples from LURE, showing distinct clusters that reveal low-rank hallucination subspaces. (b) Qualitative comparison of captions generated by MiniGPT before and after applying our proposed Wiener based correction. The baseline model hallucinates non-existent objects such as clouds, and striped pole at the top, while our method suppresses these errors.

adaptations, or rely on external resources. We address this gap with a gradient-free and inference-efficient correction derived from a representation-space estimation framework. We model hidden states as the superposition of a truthful component and a hallucination-associated distortion, each characterized by second-order statistics estimated from paired truthful and hallucinatory generations. Under this formulation, hallucination mitigation reduces to a linear minimum mean-square error (MMSE) estimation problem.

Empirically, we observe structured geometric separation between truthful and hallucinatory representations in deeper layers. As shown in Fig. 1(a), UMAP projections reveal distinct latent regions for truthful (blue) and hallucinatory (amber) states, suggesting that hallucination manifests as anisotropic distortion in representation space.

The MMSE formulation yields a closed-form Wiener estimator whose operator depends on the covariances of truthful and hallucination components. Solving a generalized eigenvalue problem identifies spectral modes ranked by distortion-to-signal ratios, which are smoothly reweighted by Wiener gains to produce covariance-driven filtering in representation space. This continuous formulation ensures stability: small covariance estimation errors induce proportionally bounded changes in the operator.

The resulting transformation acts as a spectral equalizer in latent space. We estimate it once from a calibration set and apply it offline to the feed-forward output projections of selected deeper LLM layers. A single sharpness exponent α controls attenuation strength, suppressing hallucination-dominated modes while preserving semantically grounded directions. Because the edit is absorbed into existing weights, the architecture, parameter count, and inference cost remain unchanged.

Inspired by classical Wiener filtering in signal processing (Oppenheim, 1999), our approach interprets hallucination as structured spectral distortion and corrects it through covariance-driven spectral shaping. Evaluations on LLaVA-1.5 Liu et al. (2023), MiniGPT-4 (Zhu et al., 2023), and mPLUG-Owl2 (Ye et al., 2023b) across CHAIR and POPE benchmarks demonstrate consistent reductions in object hallucination while preserving fluency and visual grounding. A qualitative comparison in Fig. 1 (b) further illustrates Wiener's effect, showing how the same model, when filtered through our method, eliminates hallucinated objects and yields faithful, visually grounded descriptions.

2 Related Works

Vision-Language Models VLMs combine visual and textual modalities for tasks such as image captioning and VQA. They typically consist of a pre-trained vision encoder (e.g., CLIP-

ViT (Radford et al., 2021)) to extract image features, a projection module that maps these features into the LLM embedding space, and a large language model (e.g., LLaMA (Touvron et al., 2023) or Vicuna (Chiang et al., 2023)), which is usually frozen during alignment and later instruction-tuned on multimodal datasets.

Modern architectures improve modularity and efficiency through three integration strategies: early fusion (Gao et al., 2023; Zhang et al., 2023; Liu et al., 2023; Bai et al., 2023), bridging (Dai et al., 2023), and mid-fusion (Alayrac et al., 2022). Early fusion prepends projected visual tokens to the LLM input, as in LLaVA (Liu et al., 2023) and Qwen-VL2 (Wang et al., 2024). Bridging compresses visual tokens into a small set of query embeddings, used in BLIP (Dai et al., 2023) and mPLUG-Owl (Ye et al., 2023a). Mid-fusion distributes cross-attention across layers, as in Flamingo (Alayrac et al., 2022) and LLaMA 3.2-Vision (Dubey et al., 2024), enabling deeper multimodal interaction.

Object Hallucinations in VLMs Visual hallucinations in vision large language models refer to the generation of plausible but factually incorrect outputs that misrepresent the input visual content (Liu et al., 2024a; Li et al., 2023c; Kim et al., 2024), encompassing a range of distortions from subtle inaccuracies to overt fabrications. These include inventing object attributes (e.g., assigning colors, sizes, or states not depicted, such as describing a grayscale image as vibrant) or fabricating spatial relationships. In multimodal tasks like visual question answering (Lia et al., 2025; Zhu et al., 2024b;a) and image captioning (Rohrbach et al., 2018; Biten et al., 2022; Zhai et al., 2023; Li et al., 2023c), hallucinations typically appear as overly assured claims that merge observed visual elements with embedded textual knowledge, resulting in responses that appear relevant to the prompt on the surface but compromise reliability in scenarios demanding accurate visual interpretation.

To mitigate hallucinations in VLMs, methods are broadly classified into training-based (Sun et al., 2024; Jing & Du, 2024; Zhao et al., 2025a) and training-free (Huang et al., 2024a; Leng et al., 2024; Chen et al., 2024; Yang et al., 2025) categories. Training-based techniques involve retraining components to strengthen vision-language alignment, such as fine-tuning the projection module or LLM with hallucination-aware losses that reward fidelity to input visuals, often leveraging datasets augmented with paired image-text contrasts. Additional strategies include reinforcement learning from human feedback (Sun et al., 2024) tailored to multimodal outputs. Recently, Jing & Du (2024) introduced FGAIF, which provides fine-grained AI-generated feedback for hallucination supervision (object existence, attributes, and relations), enabling more precise RL-based alignment of LVLMs.

Among training-free approaches, Huang et al. (2024a) proposed OPERA, which applies an over-trust penalty and a retrospection-allocation mechanism during decoding to reduce hallucinations. Leng et al. (2024) introduced Visual Contrastive Decoding, which contrasts outputs on original versus distorted visuals to suppress object hallucinations. HALC (Chen et al., 2024) is an adaptive focal-contrast decoding strategy that corrects hallucinated tokens on the fly. Similarly, Nullu (Yang et al., 2025) identifies hallucination-related directions by performing an SVD over feature differences and then removes the top singular vectors via hard projection, fully excising those directions from the model's representation space.

In contrast, our method formulates hallucination mitigation as a linear minimum mean-square error (MMSE) estimation problem. We model hidden states as a combination of truthful signal and hallucination-induced distortion and estimate their respective covariances from calibration data. Using this second-order characterization, we compute a closed-form Wiener operator that proportionally attenuates hallucination-dominated modes based on their distortion-to-signal ratios, preserving semantically meaningful directions. This approach provides stable correction without removing any component entirely, maintaining the integrity of the model's internal representations while reducing hallucinations.

An earlier version of this work (Ali et al., 2024) proposed suppressing hallucination-related directions via low-pass spectral filter. The present work substantially extends that formulation by introducing a Wiener-filter-based representation editing framework that performs smooth, covariance-aware attenuation rather than a spectral truncation. This generalization explicitly models both signal and noise statistics, yielding a more principled and

flexible mechanism for suppressing hallucination-related components while preserving task-relevant information.

3 Method

This section introduces the VLM background and notation, analyzes the covariance structure of hallucinated representations, derives the Wiener spectral filter and its stability guarantees, and describes how the resulting operator is absorbed into the feed-forward projections of deeper transformer layers.

3.1 Background and Notation

Notation. Throughout the paper, vectors are denoted using bold lowercase letters (e.g., $\mathbf{x}, \mathbf{s}, \mathbf{n}$) and matrices using bold uppercase letters (e.g., $\mathbf{W}$). Scalars are in standard italic (e.g., d, α, λ_j). The operator norm is $\|\cdot\|_2$, and for a symmetric matrix $\mathbf{C}$, its spectrum is $\sigma(\mathbf{C})$.

Background Modern vision-language models (VLMs) process multimodal inputs through a three-stage pipeline. A vision encoder $\phi_v : \mathcal{I} \to \mathbb{R}^{n \times d_v}$ extracts n visual tokens of dimension d_v. A projection layer $\phi_p : \mathbb{R}^{d_v} \to \mathbb{R}^d$ maps these tokens into the language model embedding space of dimension d. The combined visual-text sequence is then processed by a large language model (LLM) with L transformer layers.

At each layer $\ell \in \{1, \ldots, L\}$, hidden states are updated via self-attention and feed-forward networks (FFNs). The FFN block is defined as

$$\mathrm{FFN}_\ell(\mathbf{x}) = \mathbf{W}_\ell^{\mathrm{out}} \sigma(\mathbf{W}_\ell^{\mathrm{in}} \mathbf{x} + \mathbf{b}_\ell^{\mathrm{in}}) + \mathbf{b}_\ell^{\mathrm{out}}, \tag{1}$$

where $\mathbf{W}_\ell^{\mathrm{in}} \in \mathbb{R}^{d_{\mathrm{ff}} \times d}$,$\mathbf{W}_\ell^{\mathrm{out}} \in \mathbb{R}^{d \times d_{\mathrm{ff}}}$, $\mathbf{b}_\ell^{\mathrm{in}} \in \mathbb{R}^{d_{\mathrm{ff}}}$ and $\mathbf{b}_\ell^{\mathrm{out}} \in \mathbb{R}^d$.

Let $\mathbf{x} \in \mathbb{R}^d$ denote a representation vector with covariance

$$\mathbf{\Sigma} = \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top]. \tag{2}$$

For symmetric $\mathbf{\Sigma}$, let $\mathbf{\Sigma} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ denote its eigendecomposition, where $\{\mathbf{q_j}\}$ are orthogonal spectral modes with variances $\{\lambda_j\}$. Spectral operators modify the covariance structure via functions applied to eigenvalues,

$$g(\mathbf{\Sigma}) = \mathbf{Q} g(\mathbf{\Lambda}) \mathbf{Q}^\top. \tag{3}$$

3.2 Hallucination Covariance Structure

Object hallucination corresponds to systematic discrepancies between truthful and hallucinated internal representations. Empirical visualization using UMAP (Fig. 1 (a)) reveals structured separation between truthful and hallucinated hidden states in deeper layers, indicating non-random geometric divergence.

To characterize this divergence quantitatively, we utilize a calibration dataset $\mathcal{D} = \{(I_i, c_i^+, c_i^-)\}_{i=1}^N$, where c_i^+ is a hallucinated caption generated by the model and c_i^- is a human-verified truthful caption describing the same image I_i. For each transformer layer ℓ, we extract sequence-averaged representations:

$$\mathbf{x_i^+} = \frac{1}{T_i^+} \sum_{t=1}^{T_i^+} \mathbf{h}_{\ell,t}(I_i, c_i^+), \quad \mathbf{x_i^-} = \frac{1}{T_i^-} \sum_{t=1}^{T_i^-} \mathbf{h}_{\ell,t}(I_i, c_i^-). \tag{4}$$

We model hallucination as an additive perturbation in representation space, reflecting the empirical observation that hallucinated captions produce structured shifts relative to truthful representations while preserving much of the underlying semantic signal:

$$\mathbf{x}_i^+ = \mathbf{x}_i^- + \mathbf{d}_i, \tag{5}$$

where $\mathbf{x}_i^-$ denotes the truthful representation and $\mathbf{d}_i$ captures hallucination-induced variation. We estimate two covariance matrices:

$$\mathbf{\Sigma}_T = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i^- - \boldsymbol{\mu}_T)(\mathbf{x}_i^- - \boldsymbol{\mu}_T)^\top, \qquad \mathbf{\Sigma}_H = \frac{1}{N}\sum_{i=1}^{N}\mathbf{d}_i\mathbf{d}_i^\top. \tag{6}$$

Here, $\mathbf{\Sigma}_T$ captures second-order structure of truthful representations, while $\mathbf{\Sigma}_H$ captures structured distortion induced by hallucinations. Highly anisotropic spectra (Fig. 4) indicate that hallucination manifests as directional variance amplification rather than isotropic noise.

3.3 Wiener Spectral Filtering in Latent Space

We interpret hidden representations as $\mathbf{h} = \mathbf{s} + \mathbf{n}$, where $\mathbf{s}$ is the truthful component with covariance $\mathbf{\Sigma}_T$ and $\mathbf{n}$ is hallucination-associated distortion with covariance $\mathbf{\Sigma}_H$. We assume that the truthful component $\mathbf{s}$ and hallucination distortion $\mathbf{n}$ are zero-mean and mutually uncorrelated, i.e.,

Assumption 1. *The truthful component* $\mathbf{s}$ *and hallucination distortion* $\mathbf{n}$ *are zero-mean and mutually uncorrelated:*

$$\mathbb{E}[\mathbf{s}] = \mathbf{0}, \quad \mathbb{E}[\mathbf{n}] = \mathbf{0}, \quad \mathbb{E}[\mathbf{s}\mathbf{n}^\top] = \mathbf{0}. \tag{7}$$

Under Assumption 1 and the additive model $\mathbf{h} = \mathbf{s} + \mathbf{n}$, and since $\mathbf{s}$ and $\mathbf{n}$ are uncorrelated, the covariance of the observed representation decomposes as

$$\mathrm{Cov}(\mathbf{h}) = \mathbf{\Sigma}_T + \mathbf{\Sigma}_H. \tag{8}$$

Standard LMMSE theory then gives the Wiener operator:

$$\mathbf{A}^\star = \mathbf{\Sigma}_T(\mathbf{\Sigma}_T + \mathbf{\Sigma}_H)^{-1}. \tag{9}$$

The additive model is not a literal claim that the transformer factorizes hidden states into independent semantic components; it is a paired-residual construction induced by the calibration data. For each image and layer we pair a hallucinated representation $\mathbf{x}_i^+$ with the truthful representation $\mathbf{x}_i^-$ of the same image and set $\mathbf{d}_i = \mathbf{x}_i^+ - \mathbf{x}_i^-$. After centering ($\mathbf{s}_i = \mathbf{x}_i^- - \boldsymbol{\mu}_T$, $\mathbf{n}_i = \mathbf{d}_i - \boldsymbol{\mu}_d$, $\mathbf{h}_i = \mathbf{x}_i^+ - \boldsymbol{\mu}_+$) and using $\boldsymbol{\mu}_+ = \boldsymbol{\mu}_T + \boldsymbol{\mu}_d$, we obtain $\mathbf{h}_i = \mathbf{s}_i + \mathbf{n}_i$ with zero mean. The only substantive approximation is the weak cross-covariance assumption $\mathbf{C} = \mathbb{E}[\mathbf{s}\mathbf{n}^\top] \approx \mathbf{0}$; retaining it gives the exact correlated LMMSE estimator $\mathbf{A}_C^\star = (\mathbf{\Sigma}_T + \mathbf{C})(\mathbf{\Sigma}_T + \mathbf{\Sigma}_H + \mathbf{C} + \mathbf{C}^\top)^{-1}$, and our estimator is its $\mathbf{C} \approx \mathbf{0}$ case, validated empirically in Section L.

To obtain an interpretable spectral form, we express the estimator in the eigenbasis of the hallucination covariance $\mathbf{\Sigma}_H = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$, whose spectrum empirically reveals a small set of dominant distortion directions (Appendix. D), yielding eigenmodes $\{\mathbf{q}_j\}$ with distortion variances $\{\lambda_j\}$. In this basis, the signal variance in mode j is $\tau_j^2 = \mathbf{q}_j^\top \mathbf{\Sigma}_T \mathbf{q}_j$. The distortion-to-signal ratio in mode j is $\nu_j = \lambda_j/\tau_j^2$ (with $\tau_j^2 > 0$). Mode-wise Wiener gains are

$$\gamma_j = \frac{1}{1+\nu_j} = \frac{\tau_j^2}{\tau_j^2 + \lambda_j}. \tag{10}$$

The standard Wiener gains γ_j provide the MMSE-optimal attenuation, but in practice a tunable shaping strength is desirable to compensate for covariance estimation error and distributional mismatch between the calibration set and deployment data. We therefore generalize via a sharpness exponent $\alpha > 0$:

$$\tilde{\gamma}_j = (1+\nu_j)^{-\alpha} = \left(\frac{\tau_j^2}{\tau_j^2 + \lambda_j}\right)^{\alpha}. \tag{11}$$

The resulting filter is

$$\mathbf{F}_\alpha = \mathbf{Q}\,\mathrm{diag}(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_d)\,\mathbf{Q}^\top. \tag{12}$$

The mode-wise filter $\mathbf{F}_\alpha$ is a diagonalized Wiener approximation in the eigenbasis of $\mathbf{\Sigma}_H$, not generally an exact rewriting of the full operator $\mathbf{\Sigma}_T(\mathbf{\Sigma}_T + \mathbf{\Sigma}_H)^{-1}$; it becomes exact when $\mathbf{\Sigma}_T$ and $\mathbf{\Sigma}_H$ are jointly diagonalizable. We use the eigenvectors of $\mathbf{\Sigma}_H$ to identify hallucination-dominated directions and compute per-mode truthful variance $\tau_j^2 = \mathbf{q}_j^\top \mathbf{\Sigma}_T \mathbf{q}_j$, yielding a stable approximation chosen for interpretability and numerical robustness.

Wiener filtering shapes modes according to their distortion-to-signal ratio. This continuity ensures stability: small perturbations in the covariance estimates induce proportionally bounded changes in $\mathbf{F}_\alpha$. **Stability of Spectral Shaping** Since the spectral filter is constructed from empirical covariance estimates, it is important that the resulting operator remains stable under estimation noise. The following lemma shows that operator-Lipschitz spectral maps guarantee this property, ensuring that small covariance perturbations do not produce large changes in the resulting representation operator.

Lemma 1. *Let $\mathbf{C}, \widehat{\mathbf{C}} \in \mathbb{R}^{d\times d}$ be symmetric, and let g be a scalar function defined on an interval $I \subset \mathbb{R}$ that contains $\sigma(\mathbf{C}) \cup \sigma(\widehat{\mathbf{C}})$. Assume g is* operator-Lipschitz *on I, i.e., there exists $K > 0$ such that for all symmetric $\mathbf{A}, \mathbf{B}$ with $\sigma(\mathbf{A}), \sigma(\mathbf{B}) \subset I$,*

$$\|g(\mathbf{A}) - g(\mathbf{B})\|_2 \le K \, \|\mathbf{A} - \mathbf{B}\|_2. \tag{13}$$

Then g is stable in the sense that

$$\|g(\widehat{\mathbf{C}}) - g(\mathbf{C})\|_2 \le K \, \|\widehat{\mathbf{C}} - \mathbf{C}\|_2. \tag{14}$$

Proof is provided in Appendix. B.

In our implementation the gains satisfy $0 \le f_j \le 1$, so the edit is non-expansive in the hallucination eigenbasis and amplifies no spectral mode. Rather than hard rank truncation, all eigenmodes are retained and attenuated continuously, reducing sensitivity to unstable eigenvector ordering. We clamp tiny negative τ_j^2 to zero and leave modes unchanged when $\tau_j^2 + \lambda_j$ is numerically zero. The sensitivity analysis in Appendix E shows performance remains stable over a broad range of α.

Interpretation. The mode-wise gains are set by the distortion-to-signal ratios derived from the covariance structure, and α acts only as a *sharpness* control that adjusts how strongly these gains are expressed without changing their ordering. The transformation therefore remains stable and model-adaptive, governed by covariance geometry rather than downstream tuning, and admits a broad operating range for α, typically on the order of 1–10^2 (Appendix E).

3.4 Weight Correction and Inference

We incorporate the filter into a selected subset of deeper layers $\mathcal{L} \subseteq \{1, \ldots, L\}$, where hallucination-induced separation is empirically most pronounced (see Section 4.2 for an ablation over layer depth), by modifying the FFN output projection $\widetilde{W}_\ell^{\text{out}} = F_\alpha \, W_\ell^{\text{out}}$, where $F_\alpha \in \mathbb{R}^{d\times d}$ acts on the output feature dimension.

This edit corresponds to applying the Wiener estimator in representation space. Because $\mathbf{F}_\alpha$ is precomputed offline and absorbed into the weights, no additional computation is introduced at inference time. The architecture, and runtime remain unchanged.

4 Experiments

Evaluation Datasets We evaluate the proposed Wiener-based correction on three complementary hallucination-centric benchmarks. (1) CHAIR (Rohrbach et al., 2018) measures hallucination frequency by comparing generated captions with MSCOCO (Lin et al., 2014) object annotations, reporting CHAIR_S (sentence-level rate) and CHAIR_I (object-level rate); lower is better, over 500 MSCOCO validation images averaged across three seeds. (2) POPE (Li et al., 2023b) is a VQA-style yes/no benchmark probing object presence, with

Prompt: How does this meme reflect or comment on Elon Musk's public image, personality, or actions?

mPLUG-OWL2: The meme of Elon Musk holding a dog in a spacesuit reflects or comments on his public image, personality, or actions by portraying him as a playful and eccentric individual. The image of Elon Musk holding a dog in a spacesuit is a lighthearted and humorous take on his work in the space industry and his ambitious plans for colonizing Mars. It also highlights his eccentric personality, as he is known for his unconventional ideas and actions. The meme serves as a way to entertain and engage with his fans and followers, while also showcasing his unique perspective on space exploration and technology.

+ Ours: This meme reflects or comment on Elon Musk's public image, personality, or actions by portraying him as a monkey-like creature holding a Dogecoin (a cryptocurrency) in his hands. The monkey-like creature and Dogecoin symbolize Musk's playful and sometimes controversial nature, as well as his involvement in cryptocurrency and Dogecoin-related projects. The meme's depiction of Musk as a monkey-like creature, rather than a human, further emphasizes his perceived wild and unorthodox behavior, which has been criticized by some. However, it also showcases his willingness to embrace and adapt to new technologies and trends, such as cryptocurrency and Dogecoin.

Figure 2: A qualitative case study comparing model-generated descriptions for a complex meme. We visualize the outputs from the Baseline and our method, color-coding text segments as Hallucinations and Truth.

three negative-sampling strategies (random, popular, and adversarial); we report accuracy, precision, and F1 on 500 MSCOCO images with ten queries each. (3) MME (Fu et al., 2023) is a broader diagnostic suite for multimodal reasoning and visual grounding, complementing CHAIR and POPE across object presence, attributes, and spatial relationships.

Vision-Language Models. We evaluate the proposed Wiener correction on three open-source VLMs representing different vision-language fusion paradigms. Liu et al. (2023) present LLaVA-1.5, which adopts early fusion, concatenating Radford et al. (2021) CLIP visual tokens with text embeddings before processing with Vicuna-7B, emphasizing modularity and efficiency. Zhu et al. (2023) introduce MiniGPT-4 which follows a bridging design, where Q-Former–compressed visual features are injected into LLaMA-7B, reducing token dimensionality while preserving alignment. Ye et al. (2023b) propose mPLUG-Owl2 which employs a mid-fusion strategy with modular connectors and dual-branch pathways to balance visual understanding and text generation. These architectures span early, bridged, and mid-level fusion mechanisms, enabling evaluation across diverse representation geometries.

Baselines. We compare against representative hallucination-mitigation approaches. Decoding-based strategies operate at inference time without modifying parameters: beam search (Freitag & Al-Onaizan, 2017b); Visual Contrastive Decoding (VCD) (Leng et al., 2024), which contrasts logits under original and perturbed visual inputs; OPERA (Huang et al., 2024a), which adds an over-trust penalty and a retrospection mechanism; HALC (Chen et al., 2024), an adaptive focal-contrast decoding strategy; and MARINE (Zhao et al., 2025b), which refines generation using visual-text alignment signals. Weight-editing approaches instead modify parameters offline: Nullu (Yang et al., 2025) identifies hallucination-associated directions via eigendecomposition of feature differences and removes dominant components through orthogonal projection.

Implementation Details All experiments are conducted using the official model implementations, with the proposed Wiener edit applied *only* to the `down_proj` matrices in the FFN blocks of selected LLM layers. Hyperparameters are determined using a small held-out validation set of 100 randomly sampled MSCOCO images (disjoint from the evaluation set). We select the intervention layer range $\mathcal{L}$ and the sharpness parameter α based on performance on this validation subset, and fix the resulting configuration across all reported experiments. This procedure yields consistent, model-specific configurations. For LLaVA-1.5, we use layers 20–32 with $\alpha = 60.0$. For mPLUG-Owl2, we use layers 20–32 with $\alpha = 20.0$. For MiniGPT-4, we use layers 24–32 with $\alpha = 10.0$. These configurations are applied uniformly across CHAIR, POPE, and MME evaluations without further tuning.

For generation, we use beam search with beam size 3 on CHAIR, and greedy decoding (beam size 1) for MME. Unless otherwise specified, the maximum generation length is set to 64 tokens on CHAIR and POPE, 128 tokens for MME. We use the official evaluation scripts from HALC Chen et al. (2024) for all CHAIR-based evaluations.

Method	LLaVA-1.5			MiniGPT-4			mPLUG-Owl2		
	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	BLEU↑	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	BLEU↑	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	BLEU↑
Greedy	18.87±1.68	6.03±0.61	0.158±0.002	21.20±1.60	7.40±0.52	0.156±0.002	23.33±1.97	8.50±0.96	0.151±0.002
Beam Search Freitag & Al-Onaizan (2017b)	17.40±1.00	5.67±0.46	0.159±0.001	21.47±3.23	7.30±0.95	0.160±0.001	19.60±0.20	7.27±0.31	0.154±0.001
DoLa Chuang et al. (2024)	19.33±1.29	6.03±0.35	0.158±0.002	21.53±1.66	7.67±0.68	0.156±0.002	21.87±1.89	8.77±0.72	0.151±0.001
VCD Leng et al. (2024)	23.87±2.72	7.70±0.90	0.144±0.010	21.13±2.25	7.63±0.68	0.154±0.001	27.07±1.22	10.23±0.55	0.139±0.003
OPERA Huang et al. (2024a)	17.60±2.09	5.63±0.42	0.160±0.003	21.73±2.04	7.63±0.47	0.173±0.002	17.47±0.61	7.33±0.72	0.151±0.002
MARINE Zhao et al. (2025b)	18.27±0.76	5.90±0.36	0.155±0.002	20.27±2.42	7.27±0.75	0.152±0.001	–	–	–
HALC Chen et al. (2024)	16.13±0.99	5.23±0.47	0.159±0.002	17.67±1.40	6.57±0.84	0.175±0.002	17.27±0.58	7.10±0.61	0.153±0.001
Nullu Yang et al. (2025) (Greedy)	19.00±1.25	5.93±0.57	0.152±0.002	18.87±1.50	7.03±0.42	0.153±0.002	18.80±1.91	7.27±0.97	0.149±0.001
Nullu Yang et al. (2025) (Beam)	17.60±0.87	5.90±0.36	0.153±0.002	19.07±3.21	6.97±1.04	0.158±0.001	17.13±1.72	6.83±1.10	0.151±0.002
Ours	**14.93**±2.61	**4.70**±0.66	0.151±0.005	**16.87**±2.81	**6.13**±1.33	0.157±0.002	**15.40**±1.93	**6.07**±0.60	0.142±0.001

Table 1: Comparison of VLMs (LLaVA-1.5, MiniGPT-4, and mPLUG-Owl2) on MSCOCO using strategies for hallucination mitigation. $CHAIR_S$ and $CHAIR_I$ measure sentence- and instance hallucinations (lower is better). All experiments use a maximum token limit of 64.

Setting	Model	Greedy			HALC			VCD			OPERA			Nullu			Nullu w/ beam			Ours			Ours w/ beam		
		Acc	Prec	F_1	Acc	Prec	F_1	Acc	Prec	F_1	Acc	Prec	F_1	Acc	Prec	F_1	Acc	Prec	F_1	Acc	Prec	F_1	Acc	Prec	F_1
Random	LLaVA-1.5	80.60	72.90	83.39	80.70	73.04	83.45	74.23	67.28	78.55	79.63	72.33	82.50	83.33	76.57	85.22	79.17	71.63	82.26	**85.16**	**79.28**	**86.52**	78.13	70.30	81.66
	mPLUG-Owl2	83.57	76.21	85.59	83.73	76.44	85.71	78.50	71.89	81.32	87.47	82.26	88.40	78.27	70.23	81.87	85.03	78.92	86.46	82.70	75.53	84.83	**88.20**	**85.06**	**88.70**
	MiniGPT4	75.57	68.57	79.44	78.83	86.38	76.39	65.87	62.38	70.08	77.17	71.24	79.96	82.40	79.49	83.23	83.17	84.48	82.84	**84.96**	91.20	**83.59**	82.80	**92.48**	80.85
Popular	LLaVA-1.5	71.73	64.36	77.51	72.10	64.69	77.72	69.57	62.97	75.74	73.90	66.57	78.62	75.60	68.16	79.75	70.60	63.55	76.67	**78.43**	**71.30**	**81.53**	71.13	63.85	77.13
	mPLUG-Owl2	75.80	67.97	80.13	75.97	68.14	80.23	74.80	68.08	78.75	79.97	72.85	82.67	72.03	64.48	77.82	79.13	71.92	82.08	77.43	69.80	81.08	**83.40**	**78.17**	**84.80**
	MiniGPT4	58.17	54.74	69.29	71.30	72.58	70.46	60.27	57.37	66.80	66.50	61.05	73.12	69.50	64.37	74.12	71.30	67.76	73.90	**76.73**	76.80	**76.70**	76.53	**79.56**	75.26
Advers.	LLaVA-1.5	65.10	59.17	73.62	65.47	59.45	73.81	64.83	59.36	72.79	69.37	62.64	75.81	68.33	61.79	75.21	66.13	60.02	74.04	**70.86**	**64.03**	**76.57**	65.23	59.26	73.61
	mPLUG-Owl2	68.27	61.51	75.46	68.47	61.68	75.57	70.03	63.73	75.63	72.67	65.55	77.75	65.87	59.64	74.19	72.37	65.27	77.58	69.73	62.84	76.18	**75.96**	**69.46**	**79.40**
	MiniGPT4	56.00	53.39	68.21	69.57	70.01	69.23	59.10	56.42	66.15	63.60	58.77	71.46	65.40	60.70	71.62	69.17	65.43	72.49	73.40	71.99	**74.22**	**73.76**	**73.94**	73.13

Table 2: POPE evaluation under Random, Popular, and Adversarial settings. We report Accuracy (Acc), Precision (Prec), and F_1 score. Higher is better.

4.1 Quantitative Results

CHAIR. Table 1 reports caption hallucination and caption quality on CHAIR across three VLM backbones. Our *Wiener* weight edit consistently achieves the lowest hallucination rates across all evaluated models. Specifically, it reduces $CHAIR_S$ to **14.93**, **16.87**, and **15.40** for LLaVA-1.5, MiniGPT-4, and mPLUG-Owl2, respectively, outperforming prior approaches such as HALC and OPERA. Similarly, instance-level hallucination ($CHAIR_I$) is reduced to **4.70**, **6.13**, and **6.07**, yielding the best performance across all architectures. While some decoding-based methods achieve slightly higher BLEU scores, our approach maintains competitive caption quality while consistently minimizing hallucinated objects.

POPE. Table 2 evaluates object-presence grounding under the Random, Popular, and Adversarial POPE settings. Our method consistently improves factual grounding across the evaluated architectures and splits. For LLaVA-1.5, our approach achieves the strongest results in most settings, obtaining the highest accuracy and F_1 scores in the Random, Popular, and Adversarial splits compared to both decoding-time mitigation strategies (HALC, VCD, OPERA) and representation-editing approaches such as Nullu. MiniGPT-4 exhibits similar trends, where our method significantly improves accuracy and F_1, while also achieving particularly high precision, indicating fewer false-positive object predictions. For mPLUG-Owl2, the beam-search variant of our method yields the best performance across the three splits, outperforming competing approaches in both accuracy and F_1.

MME. We evaluate our method on the 10 relevant subsets of the MME benchmark, including OCR, count, scene, artwork, color, existence, position, celebrity, landmark, and posters (code reasoning, numerical calculations, text translation, and commonsense reasoning are irrelevant). Figure 3 reports Dolan–Moré profiles, which plot the fraction of tasks for which a method achieves performance within a ratio τ of the best method, providing a holistic comparison across diverse subsets. Our approach consistently achieves the strongest profile on both mPLUG-Owl2 and LLaVA-1.5, dominating competing methods over a wide range of τ. These results demonstrate that Wiener-based spectral shaping improves general visual grounding beyond object hallucination benchmarks. Per-subset MME scores for LLaVA-1.5 are reported in Appendix F.

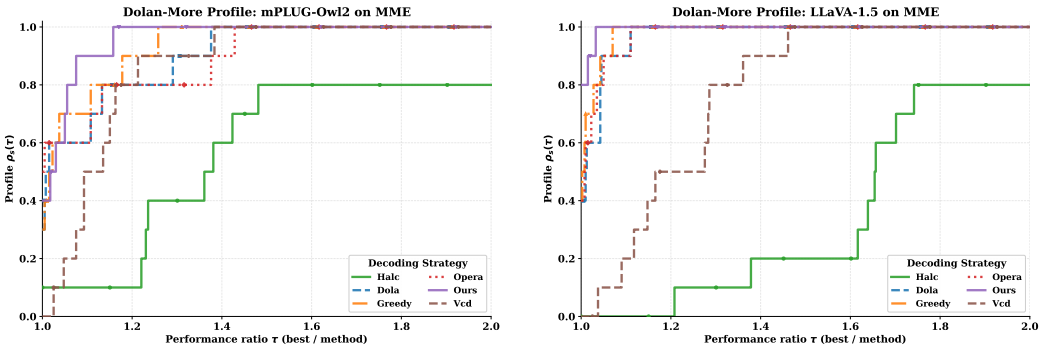


Figure 3: Dolan–Moré performance profiles on 10 MME subsets for mPLUG-Owl2 (left) and LLaVA-1.5 (right).

Additional experiments evaluating the method on Gemma3-4B-it (Appendix G), per-subset MME scores (Appendix F), the TempCompass video understanding benchmark (Appendix I), cross-domain transfer (Appendix J), and a faithfulness-versus-informativeness analysis (Appendix H) are provided in the supplementary material.

FaithDial. Table 3 reports the results on FaithDial (Dziri et al., 2022). Applying the Wiener-based representation filtering to Dream-v0-Instruct-7B[1] (Ye et al.) substantially improves both grounding quality and hallucination rates. On overlap-based metrics, the filtered model achieves consistent gains, improving BLEU by +2.21 and ROUGE-1/2/L by +2.81, +3.07, and +3.36, respectively. Grounding-oriented metrics also show large improvements, as can be seen, Token-F1 increases by 9.86 points and Q^2 by 10.56, indicating stronger alignment between generated responses and the provided knowledge. Most importantly, hallucination rates drop significantly. FaithCritic reports a reduction from 92% to 74%, while the Claude Sonnet 4.6 judge shows an even larger decrease from 93% to 50%, corresponding to a 43-point absolute reduction in unfaithfulness rate. These results suggest that the proposed representation filtering not only reduces hallucinated statements in diffusion-based language generation but also improves the factual grounding and content fidelity of the produced responses.

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	Token-F1	Q^2	FaithCritic↓
Dream	3.89	25.32	7.28	18.97	29.00	19.19	85.92
With Nullu	3.95	25.95	7.83	19.53	**35.00**	24.63	82.91
With Wiener	**4.47**	**26.22**	**8.16**	**20.45**	33.30	**26.24**	**75.39**

Table 3: Full FaithDial test set evaluation on Dream-v0-Instruct-7B.

4.2 Ablation Studies

We analyze which components of the proposed Wiener edit contribute to the observed improvements by performing controlled ablations on the CHAIR benchmark using MiniGPT-4. Each ablation isolates a specific element of the representation correction while keeping the intervention layers and evaluation protocol identical. Results are reported in Table 4.

Mean distortion subtraction. The first ablation removes only the empirical mean distortion $\mu_d = \frac{1}{N}\sum_i d_i$ from the MLP output, without applying the Wiener operator. This corresponds to a first-order correction that compensates for a global bias between truthful and hallucinated representations. While this adjustment slightly reduces hallucination metrics, the improvement remains modest. This behavior indicates that hallucination is not primarily caused by a constant offset in representation space. Instead, hallucinated

[1] https://huggingface.co/Dream-org/Dream-v0-Instruct-7B

Metric	Baseline	Mean distortion subtraction	Uniform shrinkage	layers 4–12	layers 14–24	Our (layers 24–32)
CHAIR$_S$ ↓	23.0	22.0	24.0	22.0	14.0	**13.0**
CHAIR$_I$ ↓	8.4	8.2	8.5	7.0	5.0	**4.9**
BLEU↑	0.157	0.157	0.157	0.186	0.166	0.162

Table 4: Ablation analysis on CHAIR using MiniGPT-4. Lower CHAIR$_S$ and CHAIR$_I$ indicate fewer hallucinations, while higher BLEU indicates better caption quality.

activations exhibit structured variation that cannot be captured by a single mean vector. Consequently, subtracting the mean distortion removes only a small portion of the error while leaving the dominant hallucination directions largely unchanged.

Uniform shrinkage. In the second ablation we replace the covariance-driven operator with an isotropic scaling $P = cI$, applying the same editing pipeline but without exploiting the covariance structure of the distortions. This operation uniformly contracts the representation space regardless of the direction of variation. Empirically, this naive shrinkage slightly degrades performance relative to the baseline. The degradation suggests that hallucination-related variance is highly anisotropic: certain directions are strongly associated with hallucinations, while others encode useful semantic or visual information. Uniform scaling therefore suppresses informative directions together with hallucination-dominated ones, leading to a loss of useful signal

Layer-wise Wiener editing. We analyze the effect of applying the Wiener edit across different layer ranges. As shown in Table 4, editing early layers (4–12) yields only marginal improvements, indicating that low-level representations are less directly associated with hallucination structure. In contrast, intervening in mid-level layers (14–24) leads to a substantial reduction in both CHAIR$_S$ and CHAIR$_I$, suggesting that hallucination-related distortions emerge at semantic abstraction stages. The strongest gains are obtained when editing deeper layers (24–32), where the method achieves the lowest hallucination rates, confirming that distortion is most pronounced near the output stage. These results demonstrate that Wiener-based spectral shaping is most effective when applied to high-level representations, where it aligns generation with visual evidence while preserving fluency. Additional ablations of the Wiener Sharpness parameter α could be found in Appendix. E.

Assumption validation. We empirically verify the weak cross-covariance assumption underlying our estimator: image-aligned truthful/hallucinated pairs yield low covariance-additivity error and high spectral concentration, both of which collapse under a random-pair control (Appendix L).

5 Conclusion

This work addresses object hallucinations in vision-language models by analyzing the covariance structure of hidden representations and introducing a Wiener-based representation filtering framework. Hallucination mitigation is cast as a linear estimation problem in representation space, yielding a training-free, post-hoc correction that attenuates hallucination-dominated spectral modes from a lightweight one-time calibration (forward passes and empirical second-order statistics) with no gradient updates. The correction requires no architectural changes and is absorbed directly into existing weight matrices without increasing inference cost, while consistently reducing object hallucination across multiple VLMs and preserving caption quality and fluency.

Acknowlegments

This work was supported by a Tel Aviv University Center for AI and Data Science (TAD) grant and by Len Blavatnik and the Blavatnik Family foundation. This research was also supported by the Ministry of Innovation, Science & Technology, Israel (1001576154) and the Michael J. Fox Foundation (MJFF-022407). The contribution of Ameen Ali is part of a PhD thesis research conducted at Tel Aviv University.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Ameen Ali, Tamim Zoabi, Lidor Brami, and Lior Wolf. Suppressing VLM hallucinations with spectral representation filtering. *arXiv preprint arXiv:2409.09947*, 2024.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1381–1390, 2022.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

Zhaorun Chen et al. HALC: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv*, 2024.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, 2024.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo Ponti, and Siva Reddy. Faithdial: A faithful benchmark for information-seeking dialogue. *arXiv preprint, arXiv:2204.10757*, 2022. URL https://arxiv.org/abs/2204.10757.

Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In Thang Luong, Alexandra Birch, Graham Neubig, and Andrew Finch (eds.), *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics, 2017a.

Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. *Proceedings of the First Workshop on Neural Machine Translation*, 2017b.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv e-prints*, 2023.

Qidong Huang et al. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR'24*, 2024a.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. Visual hallucinations of multi-modal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9614–9631, 2024b.

Liqiang Jing and Xinya Du. Fgaif: Aligning large vision-language models with fine-grained ai feedback. *arXiv preprint arXiv:2404.05046*, 2024.

Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Discovering and mitigating visual biases through keyword explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11082–11092, 2024.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Junyi Li et al. HaluEval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*, 2023a.

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

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

Zeh Lia et al. Vision amplified: Visual question answering with enhanced medical images. In *MICCAI*, 2025.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024a.

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

Yuanxin Liu et al. TempCompass: Do video LLMs really understand videos? *arXiv preprint arXiv:2403.00476*, 2024b.

Alan V Oppenheim. *Discrete-time signal processing*. Pearson Education India, 1999.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models, 2025. *URL https://arxiv.org/abs/2508.15487*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023a.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023b.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. Halle-switch: Rethinking and controlling object existence hallucinations in large vision-language models for detailed caption. *arXiv e-prints*, 2023.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *The Twelfth International Conference on Learning Representations*, 2023.

Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Baobao Chang, and Minjia Zhang. Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19677–19701, 2025a.

Linxi Zhao et al. MARINE: Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv preprint*, 2025b.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2024.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Deyao Zhu et al. Combating hallucinations in multimodal large language models. *arXiv preprint*, 2024a.

Jiawei Zhu, Yishu Liu, Huanjia Zhu, Hui Lin, Yuncheng Jiang, Zheng Zhang, and Bingzhi Chen. Combating visual question answering hallucinations via robust multi-space co-debias learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*. Association for Computing Machinery, 2024b. URL https://doi.org/10.1145/3664647.3681663.

A Full Algorithm

Algorithm 1 summarizes the proposed Wiener editing procedure. Given paired hallucinated and truthful representations at each selected layer $\ell \in \mathcal{L}$, we estimate the truthful covariance Σ_T and the distortion covariance Σ_H, compute a Wiener-shaped spectral operator F_α from their induced distortion-to-signal ratios, and absorb this operator into the FFN output projection:

$$\widetilde{W}_\ell^{\text{out}} = F_\alpha W_\ell^{\text{out}}.$$

Since the edit is computed offline and merged directly into the existing weights, the architecture and inference-time cost remain unchanged.

Algorithm 1 Wiener Spectral Shaping for Post-hoc Editing

Require: Calibration set $D = \{(I_i, c_i^+, c_i^-)\}_{i=1}^N$, selected layers $\mathcal{L}$, sharpness α
Ensure: Edited weights $\{\widetilde{W}_\ell^{\text{out}}\}_{\ell \in \mathcal{L}}$

```
for each layer ℓ ∈ L do
    Extract paired features x_i^+ ← Feat_ℓ(I_i, c_i^+),  x_i^- ← Feat_ℓ(I_i, c_i^-)
    Form distortions d_i ← x_i^+ − x_i^- and truthful mean μ_T ← (1/N) Σ_{i=1}^N x_i^-
    Σ_T ← (1/N) Σ_{i=1}^N (x_i^- − μ_T)(x_i^- − μ_T)^⊤
    Σ_H ← (1/N) Σ_{i=1}^N d_i d_i^⊤
    Eigendecompose Σ_H = QΛQ^⊤
    for each mode j = 1, ..., d do
        τ_j^2 ← q_j^⊤ Σ_T q_j,   ν_j ← λ_j / τ_j^2,   γ̃_j ← (1 + ν_j)^{−α}
    end for
    F_α ← Q diag(γ̃_1, ..., γ̃_d) Q^⊤
    W̃_ℓ^out ← F_α W_ℓ^out
end for
Replace W_ℓ^out with W̃_ℓ^out for all ℓ ∈ L
```

B Proof of Lemma 1

Proof. Define the line segment

$$\mathbf{C}_t := \mathbf{C} + t(\widehat{\mathbf{C}} - \mathbf{C}) \quad \text{for } t \in [0,1]. \tag{15}$$

Since I is convex and contains $\sigma(\mathbf{C}) \cup \sigma(\widehat{\mathbf{C}})$, we have $\sigma(\mathbf{C}_t) \subset I$ for all $t \in [0,1]$. Let

$$\Delta := \widehat{\mathbf{C}} - \mathbf{C}. \tag{16}$$

Consider the matrix-valued function

$$\Phi(t) := g(\mathbf{C}_t). \tag{17}$$

By standard results on spectral functions (Daleckii–Kreĭn theorem), Φ is differentiable and its derivative is the Fréchet derivative of g at $\mathbf{C}_t$ applied to Δ:

$$\Phi'(t) = Dg(\mathbf{C}_t)[\Delta]. \tag{18}$$

Therefore, by the fundamental theorem of calculus,

$$g(\widehat{\mathbf{C}}) - g(\mathbf{C}) = \Phi(1) - \Phi(0) = \int_0^1 Dg(\mathbf{C}_t)[\Delta]\, dt. \tag{19}$$

Taking the operator norm and using the triangle inequality gives

$$\|g(\widehat{\mathbf{C}}) - g(\mathbf{C})\|_2 \le \int_0^1 \|Dg(\mathbf{C}_t)[\Delta]\|_2\, dt \le \left(\sup_{t \in [0,1]} \|Dg(\mathbf{C}_t)\|_{\text{op}} \right) \|\Delta\|_2, \tag{20}$$

where

$$\|Dg(\mathbf{C}_t)\|_{\mathrm{op}} := \sup_{\|X\|_2=1} \|Dg(\mathbf{C}_t)[X]\|_2 \tag{21}$$

is the induced operator norm of the Fréchet derivative.

It remains to bound $\|Dg(\mathbf{C}_t)\|_{\mathrm{op}}$ uniformly by the operator-Lipschitz constant K. Fix any symmetric $\mathbf{A}$ with $\sigma(\mathbf{A}) \subset I$ and any symmetric direction $\mathbf{X}$. For $\varepsilon \neq 0$ small enough, $\sigma(\mathbf{A} + \varepsilon\mathbf{X}) \subset I$, and by the operator-Lipschitz assumption,

$$\|g(\mathbf{A} + \varepsilon\mathbf{X}) - g(\mathbf{A})\|_2 \leq K \, \|\varepsilon\mathbf{X}\|_2 = K|\varepsilon| \, \|\mathbf{X}\|_2. \tag{22}$$

Divide by $|\varepsilon|$ and take $\varepsilon \to 0$. By the definition of the Fréchet derivative,

$$\|Dg(\mathbf{A})[\mathbf{X}]\|_2 = \lim_{\varepsilon \to 0} \left\| \frac{g(\mathbf{A} + \varepsilon\mathbf{X}) - g(\mathbf{A})}{\varepsilon} \right\|_2 \leq K\|\mathbf{X}\|_2. \tag{23}$$

Taking the supremum over $\|\mathbf{X}\|_2 = 1$ yields $\|Dg(\mathbf{A})\|_{\mathrm{op}} \leq K$. Applying this with $\mathbf{A} = \mathbf{C}_t$ for all $t \in [0, 1]$ gives

$$\sup_{t \in [0,1]} \|Dg(\mathbf{C}_t)\|_{\mathrm{op}} \leq K. \tag{24}$$

Substituting into the earlier bound finishes the proof:

$$\|g(\widehat{\mathbf{C}}) - g(\mathbf{C})\|_2 \leq K \, \|\widehat{\mathbf{C}} - \mathbf{C}\|_2. \tag{25}$$

□

C Covariance Estimation Details

We use empirical covariance estimates with minimal numerical stabilization. The hallucination covariance $\mathbf{\Sigma}_H$ is computed as the sample second moment of the paired hallucination-truth residuals $\mathbf{\Delta}_i = \mathbf{h}_i^+ - \mathbf{h}_i^-$, while $\mathbf{\Sigma}_T$ is computed as the sample covariance of mean-centered truthful activations. The main editor does not use Ledoit-Wolf shrinkage, rank truncation, or an explicit ridge term on $\mathbf{\Sigma}_H$ or $\mathbf{\Sigma}_T$.

The stabilizing steps are numerical rather than statistical regularization: we clamp tiny negative estimates of τ_j^2 to zero, guard near-zero denominators in the Wiener gain by setting the corresponding gain to 1, and compute the eigendecomposition and edit matrix in float32 before casting back to the model dtype. A small ridge term ($\epsilon = 10^{-20}$) is used only in optional diagnostic scripts that explicitly solve inverse systems for analysis, it is not used in the main weight edit.

In our experiments, we estimate the covariance statistics using the LURE dataset with approximately 3k paired truthful/hallucinated samples per model. Empirically, stable performance is achieved once the dominant hallucination covariance structure is sufficiently captured, consistent with the heavy-tailed low-rank spectrum shown in Appendix D.

D Hallucination Spectrum

Hallucination covariance spectrum. Figure 4 visualizes the eigenvalue spectrum of the hallucination covariance matrix $\Sigma_H = Q\Lambda Q^\top$ for three representative vision–language models: LLaVA-7B, MiniGPT-4, and mPLUG-Owl2. Across all architectures, the spectrum exhibits a characteristic heavy-tailed structure, with a small number of dominant eigenvalues followed by a long, smoothly decaying tail. The leading eigenvalues correspond to high-variance directions in representation space that capture the majority of hallucination-related variation. In contrast, the remaining eigenmodes contribute progressively smaller variance and behave similarly to noise-like fluctuations.

This spectral structure suggests that hallucination behavior is largely concentrated within a low-dimensional subspace of the model representation. Despite architectural differences between the evaluated VLMs, the qualitative shape of the spectrum remains consistent, indicating that this phenomenon is not model-specific but rather a general property of multimodal language models. Motivated by this spectral structure, we view hallucination as a form of structured distortion distributed across latent modes. This perspective suggests operating directly in the spectral domain, where each mode can be treated according to its distortion-to-signal characteristics. We therefore adopt a Wiener-inspired formulation, in which the representation is shaped through a covariance-driven spectral operator that balances fidelity to the underlying signal with robustness to distortion. This yields a principled mechanism for preserving semantic consistency and fluency while improving alignment with the visual input.

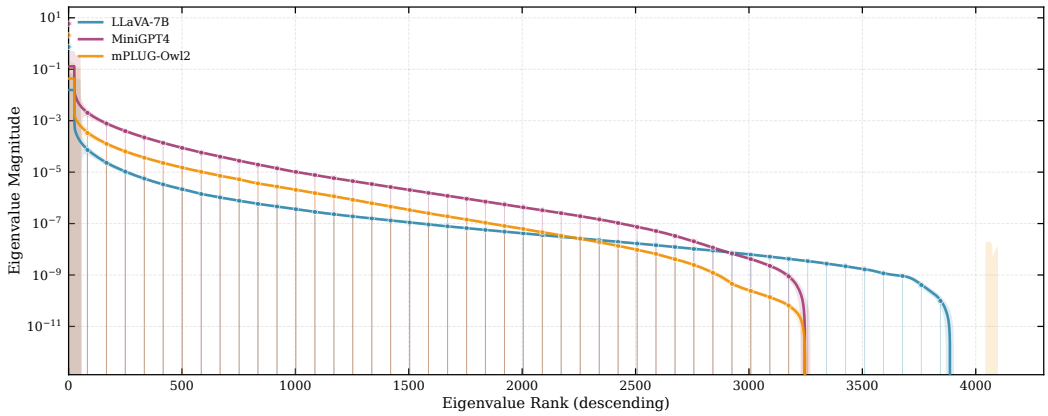


Figure 4: Hallucination spectrum (eigenvalues of $\Sigma_H = Q\Lambda Q^\top$) for three vision–language models. The curves show a small set of high-variance "spikes" followed by a long decaying tail, indicating that hallucination behavior is concentrated in a low-dimensional set of dominant eigenmodes, with the remaining spectrum reflecting noise-like variation.

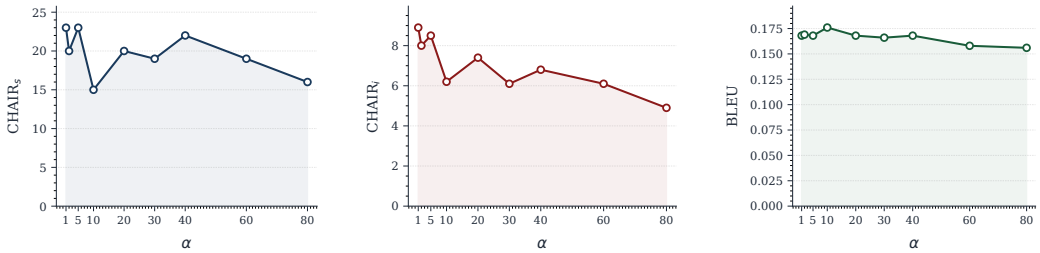
E Sensitivity of the Wiener Sharpness Parameter α

We analyze the sensitivity of the proposed Wiener edit to the sharpness parameter α. In our formulation, α does not determine which spectral modes are emphasized or de-emphasized, those ratios are determined by the Wiener gains derived from the signal and distortion covariances. Instead, α controls the *sharpness* with which these mode-wise gains are expressed, effectively adjusting how strongly the Wiener shaping is applied.

To evaluate robustness, we vary α over a wide range while keeping all other components fixed. Figure 5 reports results on 100 randomly sampled MSCOCO images using LLaVA-1.5. Panels (a) and (b) show the resulting CHAIR_S and CHAIR_I metrics, while panel (c) reports BLEU score.

Across the explored range, the method exhibits stable behavior: hallucination metrics improve consistently for moderate and large α, while BLEU remains largely unchanged. Importantly, performance does not depend on a narrowly tuned value, rather, a broad interval of α yields similar improvements. This behavior reflects the Wiener formulation itself: the relative weighting of spectral modes is governed by the covariance geometry, while α simply modulates the sharpness of this transformation.

Overall, the results indicate that the method is robust to the choice of α, with strong performance across a wide range of values. In practice, we find that values in the range $\alpha \in [10, 80]$ provide a good balance between improved grounding (lower CHAIR metrics) and stable caption quality.



(a) CHAIR_S (b) CHAIR_I (c) BLEU

Figure 5: Sensitivity of the Wiener sharpness parameter α on LLaVA-1.5 evaluated on 100 MSCOCO validation images. Lower CHAIR values indicate fewer hallucinations, while higher BLEU indicates better caption quality.

F MME Per-Subset Scores

Table 5 reports the actual per-subset MME scores for LLaVA-1.5 across all evaluated methods. Our approach achieves the highest mean score of 148.5, outperforming all competing methods including Greedy (147.0) and OPERA (145.6). Notable improvements are observed in Count (+10 over Greedy), Artwork (+5), Celebrity (+3), and Posters (+2), while maintaining competitive or identical performance on tasks such as OCR, Existence, and Position.

Method	OCR	Count	Scene	Artwork	Color	Exist.	Pos.	Celeb.	Land.	Poster	Mean
DoLa	132	138	155	125	**170**	**188**	**128**	130	150	138	145.4
Greedy	132	143	155	125	**170**	**188**	130	132	**152**	143	147.0
Halc	80	75	112	107	100	115	78	76	95	72	91.0
Opera	132	138	155	125	**170**	**188**	**128**	130	150	140	145.6
VCD	98	105	145	110	132	170	100	130	130	110	123.0
Ours	132	**153**	**157**	**130**	165	**188**	**128**	**135**	**152**	**145**	**148.5**

Table 5: MME per-subset scores for LLaVA-1.5 across all evaluated methods.

G Generalization to Stronger Models: Gemma3

To demonstrate effectiveness on more recent and stronger VLM backbones, we evaluate the proposed Wiener filtering on Gemma3-4B-it, which was released well after the models in our original evaluation. For this experiment, our method edits layers 12–26 with $\alpha = 10.0$. For Nullu, we report its best validation configuration with top_k $= 1$. Results are reported in Table 6. The proposed Wiener filter reduces CHAIR_S from 0.357 to 0.261 and CHAIR_I from 0.080 to 0.073, outperforming Nullu across all hallucination metrics while maintaining competitive caption quality (BLEU). This corresponds to a relative reduction of approximately 27% in sentence-level hallucination rate over the baseline, and roughly 16% over Nullu.

Model	$\text{CHAIR}_S\downarrow$	$\text{CHAIR}_I\downarrow$	BLEU↑	Avg obj.	Avg len.
Gemma3	0.357	0.080	**0.040**	8.21	184.2
Gemma3 + Nullu	0.311	0.077	0.038	7.37	180.1
Gemma3 + Wiener	**0.261**	**0.073**	0.037	7.11	180.2

Table 6: CHAIR evaluation on Gemma3-4B-it. The Wiener filter achieves the lowest hallucination rates while preserving caption quality.

H Faithfulness versus Informativeness

An important question is whether hallucination reduction arises from improved visual grounding or simply from more conservative generation. As shown in Table 6, the baseline Gemma3 model produces an average of 8.21 object mentions and 184.2 generated tokens per caption. After applying Wiener filtering, the average object count decreases to 7.11 and generation length to 180.2 tokens. While both editing methods slightly reduce the number of mentioned objects, the reduction is relatively small (approximately 13% in object count and 2% in generation length), whereas the improvement in hallucination metrics is substantially larger. These results suggest that the method improves the precision of object mentions by selectively suppressing hallucination-dominated representation directions while largely preserving overall generation length and descriptive content.

I Video Understanding: TempCompass

To evaluate whether the proposed representation editing framework generalizes beyond image-based hallucination, we tested on TempCompass (Liu et al., 2024b), a challenging

benchmark that measures temporal reasoning through Caption Matching (CM), Captioning (Cap), Multiple Choice QA (MCQ), and Yes/No QA (Y/N) across Action, Attribute Change, Direction, Order, and Speed dimensions. Despite being calibrated using image-based hallucination supervision, the proposed method consistently improves performance across nearly all TempCompass tasks and dimensions (Table 7). Average gains are observed from 54.2→58.5 on CM, 14.5→23.4 on Captioning, 37.1→38.9 on MCQ, and 18.4→24.0 on Y/N QA. These results suggest that the learned spectral structure captures general multimodal representation distortions that extend to temporal video reasoning. For this evaluation, we used editing layers 12–34 with $\alpha = 40.0$, without task-specific retuning.

	CM		**Caption**		**MCQ**		**Yes/No**	
Dimension	Base	Ours	Base	Ours	Base	Ours	Base	Ours
Action	58.6	**63.6**	14.1	**21.7**	46.4	**53.6**	14.0	**17.6**
Attr. Change	56.9	**59.7**	10.9	**22.4**	34.7	**36.1**	27.5	**33.0**
Direction	48.3	**54.1**	17.0	**26.8**	33.4	**34.9**	16.2	**21.7**
Order	56.3	**60.0**	22.5	**31.5**	**37.1**	31.8	25.9	**37.4**
Speed	51.5	**58.5**	7.5	**13.9**	33.1	**36.9**	12.3	**16.2**
Avg	54.2	**58.5**	14.5	**23.4**	37.1	**38.9**	18.4	**24.0**

Table 7: TempCompass evaluation on Gemma3-4B-it. The proposed method improves performance across nearly all temporal reasoning dimensions and task formats despite being calibrated on image-based hallucination data.

J Cross-Domain Transfer

To evaluate whether the learned hallucination covariance transfers across domains without retuning, we learned the covariance structure from one domain and evaluated on another. Specifically, we learned from LURE and evaluated on TruthfulQA (Lin et al., 2022), and conversely learned from HaluEval (Li et al., 2023a) and evaluated on CHAIR. In both directions, the method reduced visual hallucinations and improved textual truthfulness (Table 8), suggesting that the learned spectral structure captures hallucination-related model biases that are not limited to a single dataset or modality. All experiments were conducted on Gemma3-4B-it with layers 12–34 and $\alpha = 40.0$.

Method	**CHAIR$_S$**↓	**CHAIR$_I$**↓	**MCQ1**↑	**MCQ2**↑	**MCQ3**↑
Baseline	0.328	0.080	0.248	0.432	0.202
Nullu	0.284	0.073	0.244	0.415	0.197
Ours	**0.266**	**0.064**	**0.301**	**0.510**	**0.252**

Table 8: Cross-domain transfer on Gemma3-4B-it. CHAIR columns: filter learned from HaluEval, evaluated on CHAIR. MCQ columns: filter learned from LURE, evaluated on TruthfulQA.

K LLM Judge Prompt for FaithDial

The following prompt was used for the Claude-based LLM judge in the FaithDial evaluation:

```
``You are a fact-checker evaluating whether a language model's dialogue
response is grounded.  The model was given a knowledge passage and
dialogue history, and must respond based on that context.  GROUNDING
RULE: Every factual claim in the response must be supported by EITHER:
1.  The knowledge passage, OR 2.  The dialogue history (things already
said in the conversation).  The model is allowed to:  Paraphrase or
summarize facts from the knowledge or dialogue history, Refer back to
things already discussed in the conversation, Add pure conversational
filler with no factual content.  The model is NOT allowed to:  Add
NEW facts from its own knowledge that appear in neither the knowledge
passage nor the dialogue history, Invent specific details not in the
context, Generate broken/repetitive text. Definitions: 1. FAITHFUL: All
factual claims come from the knowledge passage or dialogue history.  2.
HALLUCINATION: The response introduces facts found in NEITHER the knowledge
NOR the dialogue history, or the output is broken/degenerate. Knowledge:
{knowledge} Dialogue History:  {history} Gold Response:  {gold_response}
Model's Response: {model_output} Output your response in valid JSON format
exactly like this: {{ ``is_hallucination'': true or false, ``category'':
``Faithful'' | ``Hallucination'', ``reasoning'': ``Brief explanation'' }}
Return ONLY the JSON object, no other text.''
```

Setting	$\rho_{\text{cross}} \downarrow$	$\delta_{\text{add}} \downarrow$	r_{16}	$\|\mathbf{A}^\star\|_F$
Paired	0	0.099	0.948	155.917
Random pair	0	1.413	0.807	346.525

Table 9: Assumption diagnostics on Gemma3-4B-it (layers 12–26). The paired setting yields low additivity error and high spectral concentration, while random pairing breaks both, confirming that the filter depends on meaningful image-aligned pairs.

L Assumption Validation

To directly validate the weak cross-covariance assumption underlying our Wiener estimator, we compute assumption diagnostics on Gemma3-4B-it using paired truthful/hallucinated LURE representations over layers 12–26. We measure: (1) normalized cross-covariance $\rho_{\text{cross}} = \|\mathbf{\Sigma}_{sn}\|_F / \sqrt{\|\mathbf{\Sigma}_T\|_F \|\mathbf{\Sigma}_H\|_F}$, (2) covariance additivity error $\delta_{\text{add}} = \|\mathbf{\Sigma}_{x^+} - (\mathbf{\Sigma}_T + \mathbf{\Sigma}_H)\|_F / \|\mathbf{\Sigma}_{x^+}\|_F$, and (3) top-$k$ residual concentration $r_k = \mathbb{E}|\mathbf{Q}_k^\top \mathbf{d}_i|_2^2 / \mathbb{E}|\mathbf{d}_i|_2^2$. We also include a random-pair control where truthful partners are shuffled across images, breaking semantic pairing. Results are shown in Table 9.

The paired setting yields a low additivity error of 0.099, and the top 16 hallucination modes explain 94.8% of the residual energy, indicating highly structured (non-isotropic) residuals. Random pairing raises the additivity error to 1.413, reduces r_{16} to 0.807, and inflates the Wiener operator norm by a relative factor of 1.222, confirming that the filter depends on meaningful image-aligned pairs rather than arbitrary covariance shrinkage.

M CHAIR Qualitative Results

Figure 6 presents representative captioning examples comparing greedy decoding, Nullu (Yang et al., 2025), and our Wiener representation edit on LLaVA-1.5. Across these examples, baseline approaches frequently introduce objects or attributes that are not supported by the visual evidence. For instance, in the beach scene, greedy decoding hallucinates *palm trees* and additional people sitting under umbrellas, while Nullu incorrectly describes the person as *carrying a surfboard*. Our method instead produces a grounded description that correctly reflects the visible elements in the image.

Similar behavior appears in the indoor dog example. Greedy decoding incorrectly states that the dog is *looking at the camera*, while Nullu misidentifies the furniture as a *couch*. In contrast, our caption correctly identifies the dog sitting on a bed and looking toward the window, consistent with the visual content.

Finally, in the train example, greedy decoding incorrectly describes the vehicle as a *toy train*, and Nullu introduces *people in the scene* that are not visible. Our method produces a more faithful description that focuses on the visible steam train and surrounding environment without introducing unsupported entities.

Overall, these examples illustrate that hallucinations often arise as spurious object insertions or attribute misidentifications. By suppressing hallucination-prone representation directions, the proposed Wiener edit produces captions that remain visually grounded while preserving descriptive quality.

N Runtime and Memory Profiling

The filter is computed offline and then absorbed into the FFN output projection weights, so it introduces no additional inference-time memory or runtime cost. To quantify the offline overhead, we profiled filter construction on Gemma3-4B-it. The computation required approximately 2.441 seconds per edited layer, with a maximum per-layer CUDA peak memory increase of 154.01 MB. Since layers are processed sequentially, this memory overhead is not

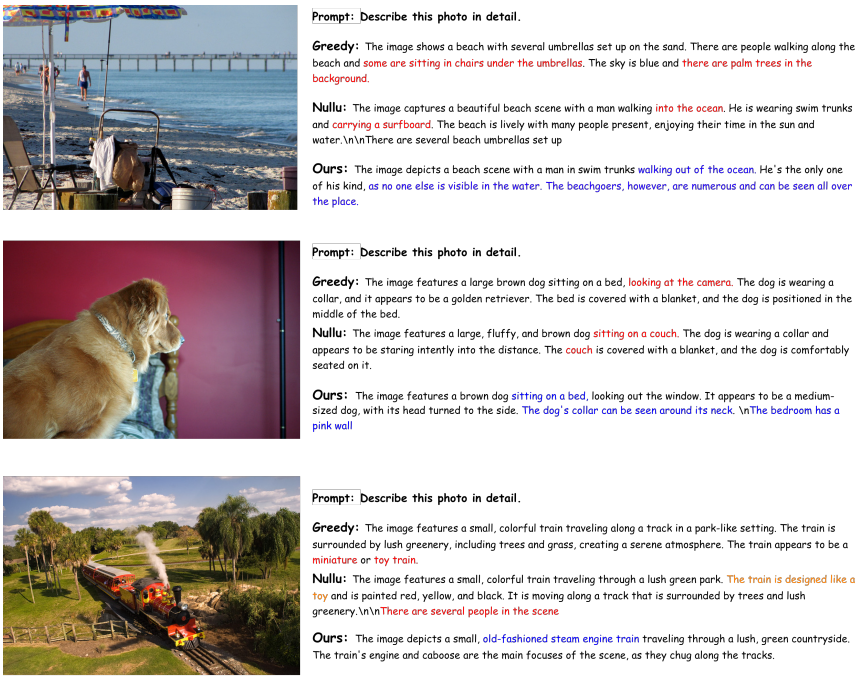


Figure 6: CHAIR captioning examples showing how Nullu and greedy decoding introduce spurious objects, while our method consistently remains visually grounded. Results are over the LLaVA 1.5 7B model.

accumulated across layers. This cost is incurred only once during the offline weight-editing stage.