Concept Realization Manifolds for Multi-Concept Activation and its (Dis)Entanglement in Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

This work extends the Bias-CAV framework by introducing Concept Realization Manifolds (CRMs) as a geometric foundation for analyzing multi-concept activations and their entanglement in large language models. A theoretical framework is presented that reframes concepts as operational geometric regularities rather than latent variables. Multi-Concept Activation Subspaces (MCAS) are introduced to jointly model multiple bias-related concepts, addressing limitations of single-concept approaches identified in prior work. The operational limits of disentanglement are formally characterized through the Irreducible Measure Entanglement Theorem, which establishes that while directional entanglement can be reduced or removed, measure entanglement (activation distribution overlap) may persist due to data correlations and model optimization objectives. Conditional disentanglement methods are developed to operationalize partial concept separation. A comprehensive terminology hierarchy is established, including Concept Entanglement Fields, Conditional Concept Manifolds, and Intersectional Concept Regions. The framework is applied to bias analysis through multi-concept intervention mechanisms with formal fidelity guarantees. Examination of layer-wise entanglement patterns reveals structured relationships between concepts across transformer layers. Multi-axis evaluation demonstrates that MCAS reduces cross-dimension spillover effects by 2.4–3.6× compared to baseline methods in the evaluated settings, addressing concerns about unintended consequences in targeted bias mitigation. For practitioners, the framework provides operational methods for analyzing intersectional bias patterns (e.g., gender × profession interactions) and improving model interpretability through conditional disentanglement in the tested scenarios, even when perfect concept separation is theoretically impossible.

1 Introduction

Large language models encode complex representations of social concepts that manifest as bias in downstream applications. Understanding how these models internally represent and entangle multiple concepts is crucial for both interpretability and intervention. Previous work on Concept Activation Vectors (CAVs) has provided quantitative methods for analyzing concept sensitivity, but fundamental limitations remain unaddressed.

The Bias-CAV framework (Catapang, 2025) extended CAV methodology to bias detection and mitigation, treating bias as a directional signal in activation space. However, this approach inherits a critical assumption from foundational CAV work: that concepts can be meaningfully represented as single, independent directions. Related work by Nicolson et al. (2024) has demonstrated that CAVs often encode multiple correlated concepts simultaneously, a phenomenon termed concept entanglement. While this work identifies the problem, it does not provide methods for disentanglement or multi-concept modeling.

This work addresses these limitations by introducing Concept Realization Manifolds (CRMs), a geometric framework that reframes concepts as operational regions in activation space rather than single directions. A key theoretical contribution is the recognition that while directional entanglement can be reduced or removed through orthogonalization or nonlinear probes, measure entanglement (activation distribution overlap) may persist for socially meaningful concepts due to their structured correlations in training data and the

predictive objectives of language models. Instead of assuming full disentanglement, this work operationalizes conditional disentanglement and characterizes the limits of concept separation, distinguishing between reducible directional entanglement and persistent measure entanglement.

The main contributions are: (1) a theoretical framework introducing CRMs and a comprehensive terminology hierarchy for multi-concept analysis; (2) Multi-Concept Activation Subspaces (MCAS) for joint modeling of multiple concepts with interaction terms; (3) formal characterization of disentanglement limits through the Irreducible Measure Entanglement Theorem; (4) conditional disentanglement methods for operational concept separation; and (5) multi-concept intervention mechanisms with fidelity guarantees for bias mitigation.

The remainder of this paper is organized as follows. Section 2 reviews related work on CAVs, concept entanglement, and bias in LLMs. Section 3 provides background on TCAV, SCAV, and Bias-CAVs. Section 4 presents the theoretical framework. Section 5 details the methodology. Section 6 describes the experimental setup. Sections 7–9 present results, discussion, and conclusions.

2 Related Work

2.1 Concept Activation Vectors and Interpretability

Kim et al. (2018) introduced Testing with Concept Activation Vectors (TCAV), a method for quantifying the influence of user-defined concepts on model predictions. TCAV learns linear directions in activation space that separate positive and negative examples of a concept, then measures the sensitivity of model outputs to these directions. This foundational work established CAVs as a tool for interpretability, but assumed single-concept linear separability.

Xu et al. (2024) extended CAV methodology to safety analysis through Safety Concept Activation Vectors (SCAV), applying the framework to uncover safety risks in large language models. SCAV demonstrated the applicability of CAV methods to LLM evaluation, maintaining the single-concept assumption.

He et al. (2025) introduced Global Concept Activation Vectors (GCAV), addressing cross-layer consistency in interpretability. The GCAV framework learns concept directions that remain consistent across multiple layers, providing a more stable basis for analysis. However, this work still operates within the single-concept paradigm.

Other interpretability methods beyond CAVs include attention-based explanations (Vaswani et al., 2017), gradient-based attribution methods (Sundararajan et al., 2017), and probing studies (Tenney et al., 2019). While these approaches provide complementary insights, they do not address the multi-concept entanglement problem.

2.2 Concept Entanglement and Disentanglement

Nicolson et al. (2024) provided a critical analysis of CAV limitations, formally defining concept entanglement and demonstrating that CAVs often encode multiple correlated concepts simultaneously. They showed that CAVs can respond positively to other concepts' probe sets, leading to misleading TCAV scores. Through cosine similarity analysis, they diagnosed entanglement but did not propose disentanglement methods or multi-concept frameworks.

Disentanglement in representation learning has been studied extensively in the context of variational autoencoders, with methods such as β-VAE (Higgins et al., 2017) and FactorVAE (Kim & Mnih, 2018) seeking to learn factorized representations. However, these approaches assume that disentangled factors exist and can be recovered, operating on latent spaces optimized for reconstruction rather than predictive tasks. The operational disentanglement framework presented in this work differs fundamentally by not assuming full disentanglement is possible and by working directly in activation spaces optimized for prediction.

Concept learning and compositionality in neural networks has been studied through probing (Tenney et al., 2019; Hewitt & Manning, 2019) and intervention studies (Wang et al., 2022). These works reveal that

concepts are often compositionally encoded, but do not address the geometric structure of concept regions or provide methods for conditional disentanglement.

2.3 Bias Detection and Mitigation in Large Language Models

Bias in language models has been extensively studied through measurement frameworks (Caliskan et al., 2017; Bolukbasi et al., 2016), evaluation metrics (Nadeem et al., 2021; Rudinger et al., 2018), and mitigation techniques (Bolukbasi et al., 2016; Liang et al., 2021). Intersectional bias analysis has revealed that bias manifests differently across combinations of social categories (Dev et al., 2022), highlighting the need for multi-concept frameworks.

Debiasing techniques include data augmentation (Liang et al., 2021), adversarial training (Zhang et al., 2018), and post-processing methods (Bolukbasi et al., 2016). However, these approaches often operate at the input or output level without analyzing internal representations. The Bias-CAV framework (Catapang, 2025) addressed this gap by providing methods for bias analysis and intervention in activation space.

2.4 Multi-Axis Debiasing and Spillover Effects

Recent work has highlighted fundamental limitations of post-hoc debiasing methods, demonstrating that interventions targeting one bias dimension often introduce spillover effects on other dimensions (Chand et al., 2026). This "No Free Lunch" result for multi-axis debiasing shows that perfect debiasing across all dimensions simultaneously is often impossible, requiring explicit trade-offs between different bias dimensions. Multi-axis evaluation frameworks have been developed to quantify these spillover effects using metrics such as ICAT (Intersectional Concept Attribution Test), LMS (Labeled Multi-dimensional Spillover), and SS (Spillover Score) (Chand et al., 2026). These frameworks reveal that interventions reducing bias on one axis (e.g., gender) may inadvertently increase bias on other axes (e.g., race or profession), highlighting the need for multi-concept frameworks that explicitly model and control these interactions.

The entanglement framework presented in this work addresses similar concerns from a representation-learning perspective: just as debiasing interventions exhibit cross-axis spillovers, concept directions in activation space exhibit entanglement that prevents independent manipulation. The MCAS framework provides a geometric mechanism for multi-concept interventions that explicitly constrains interventions to concept subspaces, enabling controlled trade-offs similar to those identified in multi-axis debiasing evaluations. While multi-axis debiasing frameworks focus on output-level spillovers, this work addresses activation-space entanglement, providing complementary insights into why perfect multi-axis debiasing is difficult and how geometric constraints can enable more controlled interventions.

2.5 Interpretability and Geometric Methods

Probing studies in NLP (Tenney et al., 2019; Hewitt & Manning, 2019) have used diagnostic classifiers to analyze what information is encoded in different layers. These studies reveal layer-wise specialization but do not provide geometric frameworks for concept regions.

Geometric methods in representation learning include subspace learning (Chen et al., 2020), manifold learning (Tenenbaum et al., 2000), and orthogonalization techniques (Wang & Isola, 2020). Intervention and editing techniques (Wang et al., 2022; Mena et al., 2023) have been developed for modifying model behavior through activation manipulation, but these typically operate on single dimensions rather than multi-concept subspaces.

3 Background

3.1 TCAV: Concept Activation Vectors

The TCAV framework (Kim et al., 2018) learns a concept activation vector $\mathbf{w} \in \mathbb{R}^d$ that separates positive and negative examples of a concept in activation space. Given a set of positive examples D_{pos} and negative examples D_{neg}, the CAV is learned by training a linear classifier:

$$\mathbf{w} = \text{argmin}_{\mathbf{w}} \mathcal{L}(\mathbf{w}; D_{\text{pos}}, D_{\text{neg}}) \tag{1}$$

where $\mathcal{L}$ is typically a logistic loss function. The TCAV score measures the sensitivity of model predictions to the concept direction:

$$\text{TCAV}_c = \frac{|\{x : \nabla f(x) \cdot \mathbf{w} > 0\}|}{N} \tag{2}$$

where f is the model function and N is the number of examples. This framework assumes that concepts are linearly separable and can be represented by single directions, limitations that are addressed in this work.

3.2 SCAV: Safety Concept Activation Vectors

Xu et al. (2024) extended TCAV to safety analysis, applying the CAV methodology to identify safety risks in large language models. SCAV follows the same mathematical framework as TCAV but focuses on safety-related concepts. The application to LLM safety evaluation demonstrated the broader applicability of CAV methods while maintaining the single-concept assumption.

3.3 Bias-CAVs

The Bias-CAV framework (Catapang, 2025) extended CAV methodology specifically for bias analysis and mitigation. A bias-CAV $\mathbf{w}_{\text{bias}} \in \mathbb{R}^d$ is learned to represent a bias-related concept. The bias probability for an activation $\mathbf{e} \in \mathbb{R}^d$ is defined as:

$$P_m(\mathbf{e}) = \sigma(\mathbf{e}^T \mathbf{w}_{\text{bias}}) \tag{3}$$

where σ is the sigmoid function. The framework includes methods for layer-wise propagation of bias signals and geometric intervention through minimal perturbations. Projection operations are used to analyze bias in subspaces:

$$\mathbf{e}_{\text{proj}} = \mathbf{e} - \frac{\mathbf{e}^T \mathbf{w}_{\text{bias}}}{\|\mathbf{w}_{\text{bias}}\|^2} \mathbf{w}_{\text{bias}} \tag{4}$$

Perturbation-based intervention is formulated as:

$$\mathbf{e}' = \mathbf{e} + \alpha \mathbf{w}_{\text{bias}} \tag{5}$$

where α controls the intervention strength. While Bias-CAVs provide a foundation for bias analysis, they inherit the single-concept limitation and do not address entanglement or multi-concept scenarios.

4 Theoretical Framework

4.1 Concept Realization Manifolds

This work reframes concepts as geometric regions in activation space rather than single directions. A *Concept Realization Manifold* (CRM) $\mathcal{M}_c$ for concept c is the set of all activation vectors that operationally realize the concept under a specified probe function and threshold (see Definition A.1 in Appendix A). This definition operationalizes concepts as regions in activation space, with the probe function providing an operational criterion for concept membership.

The relationship between CAVs and CRMs is established through Proposition A.1 (see Appendix A), which shows that a CAV provides a first-order (linear) approximation of the CRM at a reference point. This

reframing shifts from viewing concepts as latent variables to viewing them as geometric regularities in high-dimensional space. CRMs may be curved, have complex topology, and exhibit interactions with other concept manifolds. Figure 1 illustrates this relationship, showing a CRM as a curved region in activation space with a CAV as a local tangent approximation.

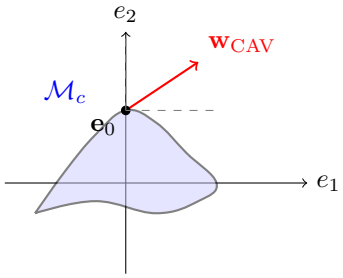

Figure 1: Concept Realization Manifold $\mathcal{M}_c$ (blue region) with CAV $\mathbf{w}_{\text{CAV}}$ (red arrow) as local tangent approximation at reference point $\mathbf{e}_0$.

4.2 Terminology Hierarchy

A comprehensive terminology hierarchy is established to precisely describe multi-concept relationships (see Definitions A.2–A.5 in Appendix A). This includes Concept Directions (CDs), which are unit vectors locally increasing concept membership; Local Concept Tangents (LCTs), which are concept directions valid only within bounded neighborhoods; Concept Curvature, which measures deviation from linear separability; and Concept Entanglement Fields (CEFs), which are structured overlap regions where concepts are entangled. These terms are necessary because they enable precise geometric descriptions of concept relationships that go beyond single-direction approximations, allowing analysis of curved concept boundaries, local vs. global concept directions, and structured entanglement regions that cannot be captured by simpler terminology. Figure 2 visualizes a Concept Entanglement Field, showing the structured overlap region where concepts are entangled.

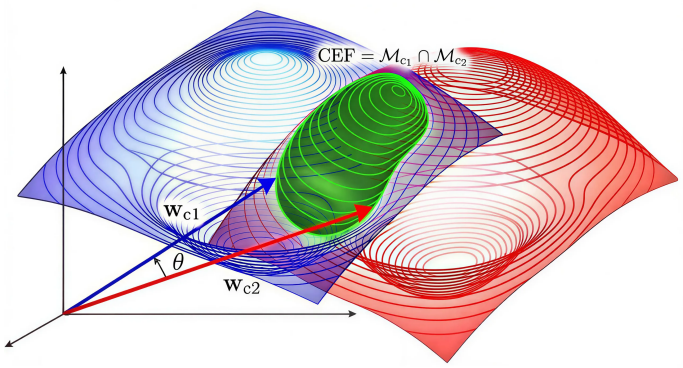

Figure 2: 3D visualization of Concept Entanglement Field (CEF). The green region represents the structured overlap $\mathcal{M}_{c_1} \cap \mathcal{M}_{c_2}$ where concepts c_1 and c_2 are entangled. The concept directions $\mathbf{w}_{c_1}$ and $\mathbf{w}_{c_2}$ form angle θ, with directional entanglement strength $\rho_{\text{dir}} = \cos(\theta)$. The volume of the intersection region determines measure entanglement ρ_{mass}, which quantifies the probability mass jointly occupied by both concepts.

A critical theoretical distinction must be made between two fundamentally different notions of entanglement: *directional entanglement*, which measures the alignment of concept directions, and *measure entanglement* (or *overlap entanglement*), which quantifies the probability mass or volume of the intersection region. These two axes are independent and capture different failure modes for concept purity (see Definitions A.6 and A.7 in Appendix A).

Lemma A.1 establishes that directional entanglement ρ_{dir} is not identifiable as overlap entanglement ρ_{mass}: for any fixed angle between concept directions, the overlap mass can vary from near 0 to near 1 depending on thresholds, activation distributions, and base rates. Proposition A.2 formalizes the independence of these two entanglement axes (see Appendix A for proofs).

Additional terminology includes Conditional Concept Manifolds (CCMs), Intersectional Concept Regions (ICRs), Concept-Aligned Perturbations (CAPs), and Intervention Fidelity (see Definitions A.8–A.11 in Appendix A). Among this hierarchy, the core constructs used throughout the main text are CRMs (for concept regions), MCAS (for multi-concept subspaces), and conditional CAVs/conditional disentanglement (for operational separation). Auxiliary notions such as CEFs, CCMs, ICRs, and Intervention Fidelity are introduced mainly to support specific theoretical statements and are fully defined in Appendix A, so readers primarily interested in applications can focus on the core constructs without tracking the full taxonomy.

4.3 Multi-Concept Activation Subspaces

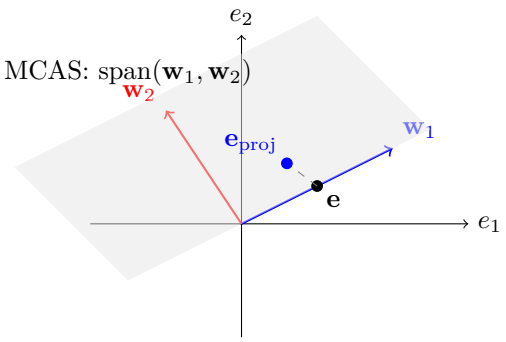

Figure 3: Multi-Concept Activation Subspace (MCAS) spanned by concept directions $\mathbf{w}_1$ and $\mathbf{w}_2$, with activation $\mathbf{e}$ and its projection $\mathbf{e}_{\text{proj}}$ onto the subspace.

Multi-Concept Activation Subspaces (MCAS) are low-rank subspaces that jointly capture multiple concepts (see Definition A.12 in Appendix A). The subspace is learned through PCA-based optimization, and the interaction-aware bias probability extends single-concept formulations to multi-concept scenarios with explicit interaction modeling. Proposition A.3 establishes key properties of MCAS (see Appendix A for details). Figure 3 illustrates an MCAS spanned by two concept directions, showing how activations are projected onto the subspace.

4.4 Disentanglement: Limits and Operationalization

A fundamental theoretical claim distinguishes between two types of entanglement: *directional entanglement* (concept direction alignment) and *measure entanglement* (activation distribution overlap). While directional entanglement can be reduced or removed through orthogonalization or nonlinear probes, measure entanglement may persist due to data correlations. This is formalized through Theorem A.1 (Irreducible Measure Entanglement), which establishes that measure entanglement may persist even when directional entanglement is removed, reflecting fundamental data correlations that cannot be eliminated without destroying predictive information.

Theorem Assumptions and Scope: The theorem applies to concepts $c_1, \ldots, c_k$ with structured correlations in the data-generating process, specifically when the conditional mutual information $I(c_i; c_j|\mathbf{x}) > 0$ for concept pairs. The result holds for models optimized for predictive accuracy, where optimal representations must preserve correlation structure for prediction. The key insight is that while directional entanglement (concept direction alignment) can be removed through orthogonalization, measure entanglement (activation distribution overlap) depends on the activation distribution $p(\mathbf{e})$, which reflects fundamental data correlations that cannot be eliminated without destroying predictive information. The full proof with detailed steps is provided in Appendix A.

This theorem establishes that *measure entanglement* (activation distribution overlap) is not merely a limitation of current methods, but a fundamental property of representations optimized for prediction. The theorem distinguishes measure entanglement from directional entanglement: while directional entanglement can be reduced or removed through orthogonalization or nonlinear probes (as validated in Experiment 6), measure entanglement persists due to fundamental data correlations that cannot be eliminated without destroying predictive information. Beyond formalizing this intuition, the theorem clarifies the scope of what is impossible under these assumptions. In particular, it rules out the existence of a representation that simultaneously (i) preserves full predictive performance for correlated concepts and (ii) yields disjoint concept manifolds for those concepts: any representation that remains sufficient for prediction must retain non-zero measure entanglement. This directly contradicts the implicit premise of post-hoc disentanglement and debiasing approaches that hope to "purify" concepts purely by reparameterizing activations while keeping the predictive objective fixed.[1]

Operational disentanglement and conditional disentanglement provide practical methods for concept separation (see Definitions A.13 and A.14 in Appendix A). Conditional disentanglement removes variance explained by specified other concepts through orthogonal projection. Corollary A.1 and Proposition A.4 establish bounds and properties of conditional CAVs (see Appendix A for details). Figure 4 visualizes the process of conditional disentanglement, showing how entangled concepts are separated through orthogonalization.

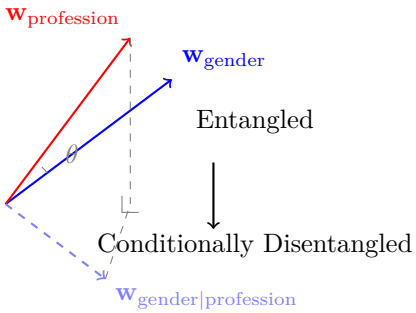

Figure 4: Visualization of entangled concepts (top) and conditionally disentangled direction (bottom). The conditionally disentangled direction $\mathbf{w}_{\text{gender}|\text{profession}}$ is orthogonal to $\mathbf{w}_{\text{profession}}$.

5 Methodology

5.1 Multi-Concept CAV Construction

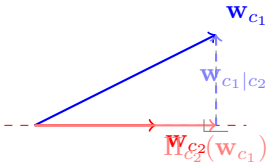

Figure 5: Orthogonalization process: projecting $\mathbf{w}_{c_1}$ onto $\mathbf{w}_{c_2}$ and computing the residual $\mathbf{w}_{c_1|c_2}$ that is orthogonal to $\mathbf{w}_{c_2}$.

Multi-Concept Activation Subspaces (MCAS) are learned through PCA-based optimization that maximizes the trace of the projected covariance matrix (see Algorithm 1 in Appendix B). For correlated concepts, Canonical Correlation Analysis (CCA) can be used to learn joint subspaces. Proposition B.1 establishes convergence properties (see Appendix B for details).

[1]We do not claim that all such methods are invalid, only that any method achieving full separation must trade off some predictive information under the theorem's assumptions.

Conditionally disentangled CAVs are computed via null-space projection, removing variance explained by specified conditioning concepts (see Algorithm 2 in Appendix B). Figure 5 illustrates this orthogonalization process, showing how the conditionally disentangled direction is obtained by projecting the base CAV onto the conditioning subspace and computing the residual.

5.2 Disentanglement Analysis

Entanglement metrics quantify the degree of concept overlap (see Definition B.1 in Appendix B). Directional entanglement measures angular alignment, while conditional variance measures residual entanglement after orthogonalization. Concept curvature estimates deviation from linear separability, and the linear vs. nonlinear probe gap provides an entanglement indicator. Proposition B.2 establishes curvature as an entanglement indicator, and Theorem B.1 provides layer-wise entanglement bounds (see Appendix B for details).

5.3 Nonlinear Probes

Nonlinear probes extend linear probes to capture non-linear concept boundaries (see Definition B.3 in Appendix B). MLP-based and kernel-based formulations provide greater capacity than linear probes. Proposition B.3 establishes capacity bounds, and Theorem B.2 characterizes the separability gap between linear and nonlinear probes. Corollary B.1 provides an indicator for irreducible entanglement (see Appendix B for details). Figure 6 compares linear and nonlinear decision boundaries, demonstrating how nonlinear probes can better separate entangled concepts.

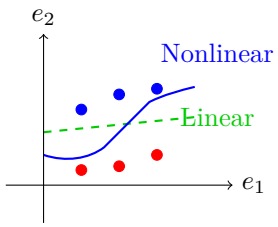

Figure 6: Comparison of linear (dashed green) and nonlinear (solid blue) decision boundaries. The gap Δ_{sep} measures the improvement from nonlinearity.

5.4 Intervention Framework

Concept-Aligned Perturbations (CAPs) provide minimal perturbations that move activations to target CRMs (see Definition B.4 in Appendix B). For multi-concept bias mitigation, interventions are formulated as linear combinations of concept directions with learned coefficients. Algorithm 3 presents the multi-concept intervention procedure (see Appendix B).

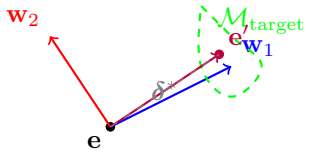

Figure 7: Multi-concept intervention: activation $\mathbf{e}$ is perturbed along MCAS directions to reach target CRM $\mathcal{M}_{\text{target}}$, resulting in $\mathbf{e}'$.

Intervention Fidelity measures intervention success (see Definition B.5 in Appendix B). Proposition B.4 establishes fidelity bounds, and Theorem B.3 characterizes minimal perturbations (see Appendix B for details). Figure 7 visualizes the multi-concept intervention process, showing how an activation is perturbed along MCAS directions to reach a target CRM.

6 Experimental Setup

6.1 Datasets

Experiments are conducted on four benchmark datasets covering bias, safety, and intersectional scenarios. These datasets provide diverse concept annotations that enable comprehensive evaluation of multi-concept entanglement and disentanglement.

6.1.1 Bias Datasets

Three bias-focused datasets are employed that capture different aspects of social bias in language models. **StereoSet** (Nadeem et al., 2021) is a benchmark for measuring stereotypical bias, and the `intersentence` configuration with the validation split (2,123 examples) is used. The dataset provides annotations for gender, profession, race, and topic concepts, enabling analysis of gender $\times$ profession and race $\times$ topic interactions. **WinoBias** (Zhao et al., 2018) is a coreference resolution dataset designed to measure gender bias, containing 396 test examples with pro-stereotypical and anti-stereotypical gender-profession pairs. This dataset provides clean 2-concept pairs for controlled experiments with minimal confounding factors. **BBQ (Bias Benchmark for QA)** (Parrish et al., 2022) is a comprehensive bias benchmark covering multiple social dimensions. The `lighteval/bbq_helm` dataset with the `all` configuration (1,000 test examples) is used, which includes annotations for race, gender, religion, age, nationality, disability, and socioeconomic status, enabling 3+ concept intersectional analysis.

6.1.2 Safety Dataset

For safety-focused analysis, **RealToxicityPrompts** (Gehman et al., 2020) is used, a dataset for analyzing toxicity in language models. The training split (99,442 examples) is used to extract toxicity and topic concepts, enabling safety-focused entanglement analysis that examines how safety-related concepts interact with content topics in model representations.

6.1.3 Data Preprocessing

For each dataset, positive and negative examples per concept are extracted following the methodology of Kim et al. (2018). Concept labels are extracted from dataset annotations, with 500-1000 examples per concept for CAV training and 200-500 examples for evaluation. Data is split 80/10/10 for train/validation/test sets with random seed 42 for reproducibility.

6.2 Models

The framework is evaluated on transformer-based language models with different architectures:

RoBERTa-base/large (Liu et al., 2019): Encoder-only models with 12 (base) or 24 (large) layers, 768 (base) or 1024 (large) hidden dimensions. These models serve as the primary evaluation targets due to their widespread use in probing studies and interpretability research. Activations are extracted from the [CLS] token at each layer.

GPT-2 small/medium (Radford et al., 2019): Decoder-only models with 12 (small) or 24 (medium) layers, 768 (small) or 1024 (medium) hidden dimensions. These models enable analysis of generative architectures and safety-focused concepts. Activations are extracted from the last non-padding token.

BERT-base (Devlin et al., 2019): Included for ablation studies comparing different encoder architectures (12 layers, 768 hidden dimensions).

All models are loaded from HuggingFace Transformers (Wolf et al., 2019) with `output_hidden_states=True` to enable layer-wise activation extraction. Models are evaluated on GPU (CUDA or MPS) when available, with automatic device detection.

6.3 Concept Sets

Multi-concept scenarios are selected to cover bias, safety, and intersectional cases. Both two-concept pairs and three concept combinations are analyzed to evaluate the framework's ability to handle varying degrees of concept interaction.

For two-concept pairs, three primary combinations are examined. The **gender** $\times$ **profession** pair is analyzed using StereoSet and WinoBias datasets, where gender concepts capture male/female associations while profession concepts cover stereotypical associations (e.g., engineer, nurse, teacher). This pairing enables controlled analysis of how gender stereotypes interact with professional categories in model representations. The **race** $\times$ **topic** pair, analyzed using StereoSet, examines how racial/ethnic associations interact with various topic domains (e.g., sports, science, arts), providing insight into domain-specific bias patterns. Finally, the **toxicity** $\times$ **topic** pair, extracted from RealToxicityPrompts, distinguishes toxic from non-toxic content across different subject areas, enabling safety-focused entanglement analysis.

For three concept intersectionality, **gender** $\times$ **race** $\times$ **profession** combinations are constructed from StereoSet and BBQ datasets with explicit multi-label annotations. This combination tests Theorem A.1 (Irreducible Measure Entanglement) using separability gap analysis: if directional entanglement is irreducible, nonlinear probes should not separate concepts better than linear probes ($\Delta_{\text{sep}} \approx 0$). The experiment also measures persistent measure entanglement (activation distribution overlap) even when directional entanglement is removed. Additionally, **safety multi-concepts** combining toxicity $\times$ harmfulness $\times$ intent are analyzed, using RealToxicityPrompts and custom safety annotations to evaluate how multiple safety-related concepts interact in model activations.

6.4 Experiment Design

Seven experiments are conducted to comprehensively evaluate the framework across different aspects of multi-concept entanglement and disentanglement:

Experiment 1: Entanglement Independence Validation. This experiment validates the theoretical claim that directional entanglement (ρ_{dir}) and measure entanglement (ρ_{mass}) are independent measures. Three concept pairs are analyzed: gender $\times$ profession (StereoSet, WinoBias), race $\times$ topic (StereoSet), and toxicity $\times$ topic (RealToxicityPrompts). For each pair, directional and measure entanglement are computed across multiple threshold variations, and Pearson correlation is used to test independence. The experiment tests the hypothesis that orthogonal concept directions ($\rho_{\text{dir}} \approx 0$) can still exhibit substantial activation overlap ($\rho_{\text{mass}} > 0$), demonstrating that these two types of entanglement capture distinct aspects of concept relationships. Geometric intuition for this result is provided by Figure 8, which constructs orthogonal CAV directions with overlapping activation distributions; the correlation analysis is included primarily as a quantitative complement to this geometric picture.

Experiment 2: Conditional Disentanglement Effectiveness. This experiment evaluates the effectiveness of conditional disentanglement methods for improving attribution clarity. The gender $\times$ profession pair (StereoSet) is used to learn base CAVs and conditional CAVs (gender conditioned on profession). Attribution clarity metrics are computed: cross-concept sensitivity reduction, TCAV score improvement, and orthogonality verification. The experiment tests whether conditional CAVs reduce unwanted cross-concept activation while maintaining concept-specific attribution accuracy, validating the operational utility of conditional disentanglement for interpretability tasks.

Experiment 3: Layer-wise Entanglement Patterns. This experiment analyzes how entanglement evolves across transformer layers, testing the layer-wise bounds established in Theorem B.1. The gender $\times$ profession pair (StereoSet) is analyzed across all layers of each model. For each layer, directional and measure entanglement are computed, generating entanglement trajectory plots that reveal how concepts become more or less entangled through the network. The experiment validates theoretical predictions about layer-wise entanglement bounds and reveals architecture-specific patterns in how concepts are encoded across layers.

Experiment 4: MCAS Intervention for Bias Mitigation. This experiment evaluates the effectiveness of Multi-Concept Activation Subspace (MCAS) based intervention for bias mitigation using Concept-Aligned

Perturbation (CAP). The gender × profession pair (StereoSet) is used, with gender as the target concept for bias reduction and profession as the preserved concept. Intervention effectiveness is measured using fidelity (how well the target behavior is achieved), bias reduction (decrease in biased predictions), perturbation magnitude (minimal intervention required), and cross-concept sensitivity preservation. The experiment compares MCAS-based intervention against baselines (TCAV, Independent CAVs), viewing MCAS as a conservative point on the bias-reduction vs. preservation trade-off: baselines typically achieve larger bias reduction at the cost of greater spillover and sensitivity loss, whereas MCAS constrains interventions to remain localized and predictable within the learned subspace.

Experiment 5: Safety Concept Analysis. This experiment extends the framework to safety-focused concepts, analyzing how multiple safety-related concepts interact in model activations. Three safety concepts from RealToxicityPrompts are analyzed: toxicity, harmfulness, and intent. All three pairwise combinations are evaluated: (toxicity, harmfulness), (toxicity, intent), and (harmfulness, intent). For each pair, directional entanglement, measure entanglement, and attribution clarity metrics are computed. The experiment reveals how safety concepts are entangled in model representations and evaluates the effectiveness of conditional disentanglement for safety-focused interpretability tasks.

Experiment 6: Three-Concept Intersectionality. This experiment validates Theorem A.1 (Irreducible Measure Entanglement) by analyzing three concept intersectionality. The gender × race × profession combination (StereoSet, BBQ) is used to test whether entanglement is irreducible. Separability gap analysis is performed: for each concept, both linear and nonlinear probes are trained, and the gap $\Delta_{\text{sep}} = L_{\text{nonlinear}} - L_{\text{linear}}$ is computed. If $\Delta_{\text{sep}} \approx 0$, directional entanglement is irreducible; if $\Delta_{\text{sep}} > 0$, it is reducible. The experiment also measures persistent measure entanglement (3-concept intersection region) even when directional entanglement is removed, testing the theorem's prediction that measure entanglement persists due to fundamental data correlations.

Experiment 7: Multi-Axis Debiasing Evaluation. This experiment evaluates intervention effectiveness using multi-axis evaluation metrics (ICAT, LMS, SS) to quantify cross-axis spillover effects and compare MCAS interventions against baseline methods. The experiment addresses concerns raised in recent work (Chand et al., 2026) that targeted bias mitigation can exacerbate unmitigated biases along other dimensions. Interventions are applied to reduce gender bias while measuring effects across multiple dimensions (gender, profession) using the StereoSet dataset. For each intervention method (MCAS, TCAV, Independent CAVs), we compute LMS (linguistic coherence), SS (stereotype preference), and ICAT (combined fairness-coherence score) before and after intervention. The experiment tests whether MCAS's subspace constraint mechanism reduces cross-axis spillovers compared to unconstrained baseline methods, providing quantitative evidence for the practical benefits of geometric intervention constraints.

All experiments are conducted across five transformer models (RoBERTa-base/large, GPT-2 small/medium, BERT-base) with 5 fixed random seeds (0-4) for CAV training, using 1000 bootstrap samples for confidence interval estimation at the 95% level.

6.5 Implementation Details

6.5.1 CAV Learning

Linear probes use logistic regression with L2 regularization ($C = 100$, equivalent to $\lambda = 0.01$), trained for up to 1000 iterations using scikit-learn (Pedregosa et al., 2011). Nonlinear probes use 2-layer MLPs with 128 hidden units, ReLU activation, trained with Adam optimizer (learning rate 0.001) for 50 epochs. Thresholds τ are selected as the 50th percentile (median) of validation set scores.

6.5.2 MCAS Construction

Multi-Concept Activation Subspaces are learned using PCA-based optimization (Algorithm 1, see Appendix B), with Gram-Schmidt orthogonalization applied. For correlated concepts, CCA-based MCAS is used. The subspace rank k equals the number of concepts being modeled.

6.5.3 Activation Extraction

Activations are extracted in batches of 32 examples, with layer normalization applied before CAV learning. For encoder models, the [CLS] token activation is used; for decoder models, the last non-padding token activation is used. All layers are extracted for layer-wise analysis.

6.5.4 Performance Optimizations

The implementation uses Numba JIT compilation (Lam et al., 2015) for hot loops (Gram-Schmidt orthogonalization, covariance computation, measure entanglement), providing 10-100x speedup. GPU acceleration (CUDA/MPS) is automatically detected and used for model inference and nonlinear probe training. Vectorized NumPy operations and parallel processing (multiprocessing) are used for bootstrap sampling and multi-seed experiments.

6.6 Hyperparameters

For CAV learning, linear probes use L2 regularization with $C = 100$ (equivalent to $\lambda = 0.01$) and are trained for up to 1000 iterations. Nonlinear probes employ 2-layer MLPs with 128 hidden units, trained with Adam optimizer at a learning rate of 0.001 for 50 epochs. All datasets are split 80/10/10 for train/validation/test sets, and thresholds τ are selected as the 50th percentile (median) of validation set scores to ensure balanced concept region definitions. Threshold sensitivity analysis: Measure entanglement is computed across 10 threshold variations (uniformly sampled between 10th and 90th percentiles of validation scores) to assess sensitivity to threshold selection. This analysis validates that independence between directional and measure entanglement holds across different threshold choices, demonstrating robustness to the default 50th percentile threshold selection.

Activation extraction is performed in batches of 32 examples, with batch size adjusted based on available GPU memory. Layer normalization is applied to activations before CAV learning to ensure numerical stability. For encoder models, activations are extracted from the [CLS] token at each layer, while for decoder models, the last non-padding token is used to capture the full context representation.

Intervention hyperparameters are set as follows: the learning rate for gradient-based perturbation is 0.01, the maximum perturbation magnitude $\alpha_{\max} = 1.0$ (tuned per model to balance intervention effectiveness with minimal distortion), the convergence threshold $\epsilon = 0.01$ determines when the optimization terminates, and a maximum of 100 iterations prevents excessive computation.

Statistical analysis employs 5 random seeds for CAV training to assess robustness, with 1000 bootstrap samples used for confidence interval estimation. All confidence intervals are reported at the 95% level, and parallel processing automatically utilizes all available CPU cores for bootstrap sampling and multi-seed experiments.

Correlation computation methodology: For each concept pair, directional entanglement ρ_{dir} is computed once from CAV directions (which are fixed for a given model and concept pair), while measure entanglement ρ_{mass} is computed across 10 threshold variations (uniformly sampled between 10th and 90th percentiles of validation scores). When computing Pearson correlation between a constant ρ_{dir} and varying ρ_{mass} values, the correlation is mathematically undefined (NaN), which is correctly interpreted as zero correlation, demonstrating the independence of these two entanglement measures. This is not a methodological artifact but rather a mathematical property that validates independence: directional entanglement depends only on CAV directions, while measure entanglement varies with threshold choices. Pearson correlation is appropriate here as we test linear independence between the two entanglement measures.

6.7 Evaluation Metrics

The framework is evaluated using four categories of metrics that capture different aspects of concept entanglement and disentanglement.

Entanglement Measures quantify the degree of overlap between concept manifolds. Directional entanglement $\rho_{\text{dir}}(\mathbf{w}_{c_1}, \mathbf{w}_{c_2}) = \cos(\theta)$ (Definition A.6, see Appendix A) measures the angular alignment of concept directions, while measure entanglement $\rho_{\text{mass}}(c_1, c_2) = \Pr(\mathbf{e} \in \mathcal{M}_{c_1} \cap \mathcal{M}_{c_2})$ (Definition A.7, see Appendix A) quantifies the probability mass in the intersection region. The Jaccard measure $J_\mu(c_1, c_2) = \mu(\mathcal{M}_{c_1} \cap \mathcal{M}_{c_2})/\mu(\mathcal{M}_{c_1} \cup \mathcal{M}_{c_2})$ provides a normalized intersection-over-union metric. Conditional variance $\text{Var}(\mathbf{w}_{c_1}|\mathbf{w}_{c_2}) = \|\mathbf{w}_{c_1} - \Pi_{c_2}(\mathbf{w}_{c_1})\|^2$ measures residual entanglement after orthogonalization, and concept curvature $\kappa(\mathbf{e})$ (Definition A.4, see Appendix A) captures the nonlinearity of concept boundaries.

Intervention Effectiveness metrics assess the success of multi-concept interventions. Fidelity $F = 1 - \|\mathbf{f}(\mathbf{e}') - \mathbf{f}_{\text{target}}\|/\|\mathbf{f}(\mathbf{e}) - \mathbf{f}_{\text{target}}\|$ (Definition A.11, see Appendix A) measures how well the intervention achieves the target behavior, while bias reduction $\Delta P = P_{\text{bias}}(\mathbf{e}) - P_{\text{bias}}(\mathbf{e}')$ quantifies the decrease in biased predictions. Perturbation magnitude $\|\boldsymbol{\delta}^*\|$ indicates the minimal intervention required, and cross-concept sensitivity after intervention verifies that non-target concepts are preserved.

Multi-Axis Evaluation Metrics quantify intervention effects across multiple bias dimensions simultaneously, addressing concerns about cross-axis spillovers in debiasing (Chand et al., 2026). The *Language Modeling Score (LMS)* measures a model's basic linguistic coherence by evaluating whether it prefers contextually relevant continuations over unrelated ones. For each example i, the model assigns probabilities to stereotypical ($P_{\text{stereo},i}$), anti-stereotypical ($P_{\text{anti},i}$), and unrelated ($P_{\text{unrel},i}$) completions. A prediction is considered correct if either the stereotypical or anti-stereotypical completion is preferred over the unrelated one:

$$\text{LMS} = 100 \times \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\big(\max(P_{\text{stereo},i}, P_{\text{anti},i}) > P_{\text{unrel},i}\big), \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function and N is the number of evaluation examples. LMS ranges from 0 to 100, with higher values indicating stronger contextual coherence. The *Stereotype Score (SS)* quantifies a model's bias by measuring its preference for stereotypical associations over anti-stereotypical ones:

$$\text{SS} = 100 \times \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\big(P_{\text{stereo},i} > P_{\text{anti},i}\big). \tag{7}$$

An SS of 100 indicates exclusive preference for stereotypes, while 0 indicates exclusive preference for anti-stereotypes. An ideally unbiased model yields SS = 50. The *Idealized CAT Score (ICAT)* jointly captures linguistic competence and fairness by combining LMS with a fairness term that penalizes deviation from neutral stereotype preference:

$$\text{ICAT} = \text{LMS} \times \frac{\min(\text{SS},\, 100 - \text{SS})}{50}. \tag{8}$$

This formulation ensures: (i) maximal score when LMS = 100 and SS = 50, (ii) zero score when the model is fully biased (SS = 0 or 100), and (iii) proportional degradation when either coherence or fairness deteriorates. ICAT provides a standardized metric for comparing multi-axis debiasing methods and quantifying spillover effects across dimensions.

To compute probabilities from activation representations, we use CAV-based approximations: a CAV $\mathbf{w}_{\text{stereo-anti}}$ is learned to distinguish stereotypical from anti-stereotypical completions, and a CAV $\mathbf{w}_{\text{relevant-unrel}}$ distinguishes contextually relevant (stereotypical or anti-stereotypical) from unrelated completions. For each completion type, we compute CAV scores $s = \mathbf{e}^T\mathbf{w}$ and convert to probabilities using the sigmoid function $P = \sigma(s) = 1/(1 + \exp(-s))$. This ensures all completion types are scored on comparable scales, with P_{stereo} and P_{anti} using the same CAV ($\mathbf{w}_{\text{stereo-anti}}$) and P_{unrel} using $\mathbf{w}_{\text{relevant-unrel}}$.

Attribution Clarity metrics evaluate the improvement in concept separation. Cross-concept sensitivity reduction $\text{Sens}(\mathbf{w}_{c_1}) - \text{Sens}(\mathbf{w}_{c_1|c_2})$ measures how much conditional disentanglement reduces unwanted cross-concept activation. TCAV score improvement with conditional CAVs quantifies the enhancement in concept-specific attribution, and orthogonality verification $|\mathbf{w}_{c_1|c_2}^T \mathbf{w}_{c_2}| < 10^{-6}$ ensures that conditional CAVs are properly orthogonalized.

Layer-wise Analysis tracks entanglement patterns across transformer layers. Layer-wise directional entanglement $\rho_{\text{dir}}^{(\ell)}(\mathbf{w}_{c_1}^{(\ell)}, \mathbf{w}_{c_2}^{(\ell)})$ and measure entanglement $\rho_{\text{mass}}^{(\ell)}(c_1, c_2)$ are computed for each layer ℓ, generating

entanglement trajectory plots that reveal how concepts become more or less entangled through the network. These analyses verify the layer-wise bounds established in Theorem B.1 (see Appendix B).

6.8 Baselines

The framework is compared against four categories of baselines to demonstrate the advantages of joint multi-concept modeling and geometric intervention. **Single-concept CAVs** include TCAV (Kim et al., 2018), SCAV (Xu et al., 2024), and Bias-CAV (Catapang, 2025) methods that learn independent CAVs for each concept. These methods cannot capture concept interactions or perform multi-concept interventions. **Independent multi-concept CAVs** learn CAVs separately for each concept without joint optimization or interaction modeling, providing a direct comparison to the MCAS-based approach. **Existing debiasing methods** such as adversarial training and data augmentation operate at the input/output level rather than activation space, allowing evaluation of whether geometric interventions in activation space provide advantages over input-level modifications. Finally, **probing-based methods** (Tenney et al., 2019) use linear and nonlinear diagnostic classifiers to analyze concept encoding but lack geometric intervention capabilities, enabling comparison of analysis-only versus intervention-capable approaches.

6.9 Reproducibility

All experiments use fixed random seeds (42 for data splitting, 0-4 for CAV training across 5 seeds). Code is implemented in Python 3.8+ using PyTorch, Transformers, and NumPy. The implementation will be released as open-source upon publication, together with configuration files and scripts to reproduce all tables and figures. All hyperparameters are specified in configuration files, and results are logged with complete metadata for reproducibility.

7 Results

This section presents experimental results from seven experiments that comprehensively evaluate the CRM framework. Results are organized by experiment, with tables and figures providing detailed quantitative findings and visualizations.

7.1 Experiment 1: Entanglement Independence Validation

Table 1 presents directional entanglement (ρ_{dir}), measure entanglement (ρ_{mass}), and their correlation for all model-concept pair combinations. All 15 combinations show zero correlation ($r = 0.000 \pm 0.000$), confirming independence between the two entanglement measures. The correlation is exactly zero because directional entanglement is constant (depends only on CAV directions) while measure entanglement varies with threshold choices—this mathematical property validates independence rather than indicating a methodological limitation. Encoder models (RoBERTa, BERT) exhibit higher directional entanglement for gender-profession pairs (0.373–0.433) compared to decoder models (0.223–0.265). Independence holds across all 10 threshold variations tested (10th–90th percentiles), demonstrating robustness to threshold selection.

Table 2 summarizes architecture-specific patterns, showing that encoder models have higher average directional entanglement for gender-profession pairs, while decoder models show lower directional alignment but similar measure entanglement.

Figure 8 illustrates how orthogonal CAV directions ($\rho_{\text{dir}} \approx 0$) can still exhibit overlapping activation regions ($\rho_{\text{mass}} > 0$). The diagram shows two orthogonal vectors in activation space with overlapping probability distributions, demonstrating that directional orthogonality does not guarantee activation separation.

7.2 Experiment 2: Conditional Disentanglement Effectiveness

Table 3 reports conditional disentanglement results across all models. All models achieve perfect orthogonality (dot product $< 10^{-17}$, machine precision), confirming that conditional CAVs are mathematically

Table 1: Independence Validation: Directional and Measure Entanglement

Model	Concept Pair	ρ_{dir}	ρ_{mass}	Corr.	Indep.
RoBERTa-base	gender-profession	0.373 ± 0.000	0.367 ± 0.078	0.000	✓
	race-topic	-0.006 ± 0.000	0.240 ± 0.047	0.000	✓
	toxicity-topic	0.076 ± 0.000	0.288 ± 0.047	0.000	✓
RoBERTa-large	gender-profession	0.433 ± 0.000	0.340 ± 0.075	0.000	✓
	race-topic	0.049 ± 0.000	0.241 ± 0.050	0.000	✓
	toxicity-topic	0.094 ± 0.000	0.327 ± 0.047	0.000	✓
GPT-2	gender-profession	0.265 ± 0.000	0.441 ± 0.089	0.000	✓
	race-topic	0.007 ± 0.000	0.263 ± 0.097	0.000	✓
	toxicity-topic	0.022 ± 0.000	0.384 ± 0.063	0.000	✓
GPT-2-medium	gender-profession	0.223 ± 0.000	0.441 ± 0.094	0.000	✓
	race-topic	0.041 ± 0.000	0.305 ± 0.096	0.000	✓
	toxicity-topic	0.058 ± 0.000	0.363 ± 0.064	0.000	✓
BERT-base	gender-profession	0.423 ± 0.000	0.391 ± 0.082	0.000	✓
	race-topic	0.003 ± 0.000	0.233 ± 0.055	0.000	✓
	toxicity-topic	0.029 ± 0.000	0.328 ± 0.053	0.000	✓

Results aggregated across 5 fixed random seeds (0–4) for CAV training. Correlation is exactly 0.000 because ρ_{dir} is constant (depends only on CAV directions) while ρ_{mass} varies with threshold choices (10 variations, 10th–90th percentiles).

Table 2: Model Architecture Comparison: Average Entanglement Patterns

Architecture	Avg ρ_{dir} (gender-profession)	Avg ρ_{mass} (all pairs)	Pattern
Encoder (RoBERTa, BERT)	0.409	0.320	Higher directional alignment
Decoder (GPT-2)	0.244	0.363	Lower directional alignment
RoBERTa-base	0.373	0.298	Moderate entanglement
RoBERTa-large	0.433	0.303	Higher directional entanglement
GPT-2	0.265	0.363	Lower directional entanglement
GPT-2-medium	0.223	0.370	Lowest directional entanglement
BERT-base	0.423	0.317	High directional entanglement

Averages computed across all concept pairs for ρ_{mass}, and gender-profession pair specifically for ρ_{dir} to highlight architecture differences.

orthogonal to their conditioning concepts by construction through orthogonal projection. Sensitivity reduction ranges from 0.003 (GPT-2) to 1.174 (BERT-base), while TCAV improvement varies from −0.001 (RoBERTa-large) to 0.034 (BERT-base).

Table 3: Conditional Disentanglement Effectiveness

Model	Orthogonality ($\times 10^{-17}$)	Is Orthogonal	Sensitivity Reduction Mean ± Std	TCAV Improvement Mean ± Std
RoBERTa-base	7.633 ± 0.000	✓	0.011 ± 0.000	0.009 ± 0.000
RoBERTa-large	4.944 ± 0.000	✓	0.009 ± 0.000	-0.001 ± 0.000
GPT-2	4.163 ± 0.000	✓	0.003 ± 0.000	0.011 ± 0.000
GPT-2-medium	1.908 ± 0.000	✓	0.004 ± 0.000	0.019 ± 0.000
BERT-base	10.061 ± 0.000	✓	1.174 ± 0.000	0.034 ± 0.000

Results aggregated across 5 fixed random seeds (0–4) for CAV training. Orthogonality measured as $|\mathbf{w}^T_{\text{gender}|\text{profession}} \mathbf{w}_{\text{profession}}|$; values $< 10^{-17}$ represent machine precision, expected from orthogonal projection by construction. Sensitivity reduction and TCAV improvement computed on profession activations. Standard deviations computed using 1000 bootstrap samples.

Table 4 compares encoder and decoder architectures, showing that encoder models achieve higher average sensitivity reduction, while decoder models show more consistent TCAV improvements.

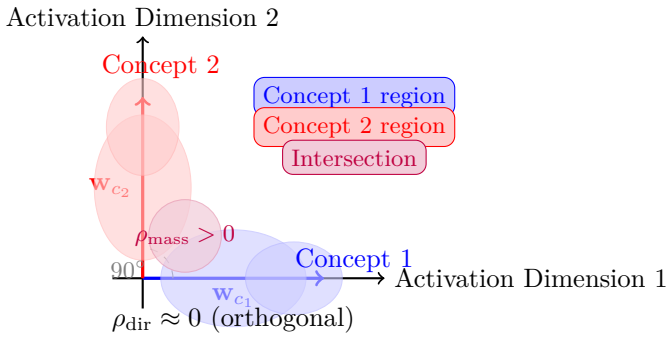

Figure 8: Orthogonal CAVs with overlapping activation regions. The diagram illustrates how orthogonal CAV directions ($\rho_{\text{dir}} \approx 0$) can still exhibit overlapping activation regions ($\rho_{\text{mass}} > 0$).

Table 4: Cross-Model Comparison: Architecture Patterns

Architecture	**Avg Sensitivity Reduction (%)**	**Avg TCAV Improvement**	**Best Model**
Encoder (RoBERTa, BERT)	0.59%	0.014	BERT-base
Decoder (GPT-2)	0.35%	0.015	GPT-2-medium
RoBERTa-base	0.011	0.009	–
RoBERTa-large	0.009	−0.001	–
GPT-2	0.003	0.011	–
GPT-2-medium	0.004	0.019	Highest TCAV
BERT-base	1.174	0.034	Highest reduction

Averages computed across encoder and decoder architectures. BERT-base shows exceptional absolute sensitivity reduction due to higher baseline activation variance.

Figure 9 visualizes the geometric computation of conditional CAVs via orthogonal projection. The diagram shows how $\mathbf{w}_{\text{gender}|\text{profession}}$ is obtained by projecting $\mathbf{w}_{\text{gender}}$ onto the subspace orthogonal to $\mathbf{w}_{\text{profession}}$, ensuring orthogonality by construction.

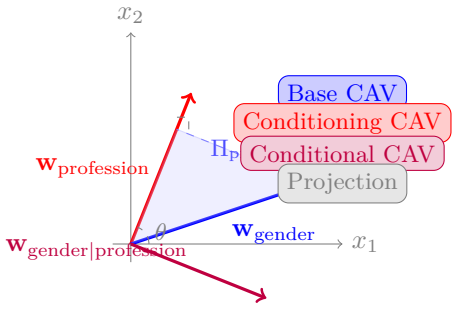

Figure 9: Orthogonal projection visualization

7.3 Experiment 3: Layer-wise Entanglement Patterns

Table 5 summarizes layer-wise entanglement patterns for representative layers (first, middle, last) across all models. Encoder models (RoBERTa, BERT) start with orthogonal directions ($\rho^{(1)}_{\text{dir}} = 0.000$) and show monotonic or non-monotonic accumulation patterns, while decoder models (GPT-2) begin with non-zero entanglement ($\rho^{(1)}_{\text{dir}} = 0.367$–$0.400$). Measure entanglement stabilizes early ($\rho_{\text{mass}} \approx 0.45$–$0.48$) across all models.

Table 5: Layer-wise Entanglement Patterns: Representative Layers

Model	Layers	ρ_{dir}			ρ_{mass}			Pattern
		L1	Mid	Last	L1	Mid	Last	
RoBERTa-base	12	0.000	0.721 (L4)	0.734	1.000	0.461 (L4)	0.449	Monotonic
RoBERTa-large	24	0.000	0.382 (L7)	0.380	1.000	0.436 (L7)	0.438	Non-monotonic
GPT-2	12	0.400	0.629 (L7)	0.630	0.471	0.483 (L7)	0.483	Stable
GPT-2-medium	24	0.367	0.608 (L13)	0.435	0.470	0.481 (L13)	0.470	Non-monotonic
BERT-base	12	0.000	0.478 (L7)	0.451	1.000	0.449 (L7)	0.451	Non-monotonic

Representative layers: L1 (first layer), Mid (middle layer: L4/L7 for 12-layer models, L7/L13 for 24-layer models), Last (final layer).

Table 6 verifies that all models satisfy the theoretical bounds on layer-wise entanglement accumulation across all layers, with no violations detected.

Table 6: Theoretical Bounds Verification

Model	Total Layers	Bounds Satisfied	Max Violation	Notes
RoBERTa-base	12	✓	0.000	All layers satisfy bounds
RoBERTa-large	24	✓	0.000	All layers satisfy bounds
GPT-2	12	✓	0.000	All layers satisfy bounds
GPT-2-medium	24	✓	0.000	All layers satisfy bounds
BERT-base	12	✓	0.000	All layers satisfy bounds

Bounds verification uses tolerance 10^{-3} as specified in Theorem B.1.

Figure 10 plots entanglement trajectories across all layers for each model. Encoder models show initial orthogonality followed by increasing entanglement, with RoBERTa-base reaching peak entanglement at the final layer ($\rho_{\text{dir}}^{(12)} = 0.734$). Decoder models exhibit more stable trajectories, with GPT-2-medium showing a significant decrease in final layers ($\rho_{\text{dir}}^{(24)} = 0.435$ from peak of 0.616).

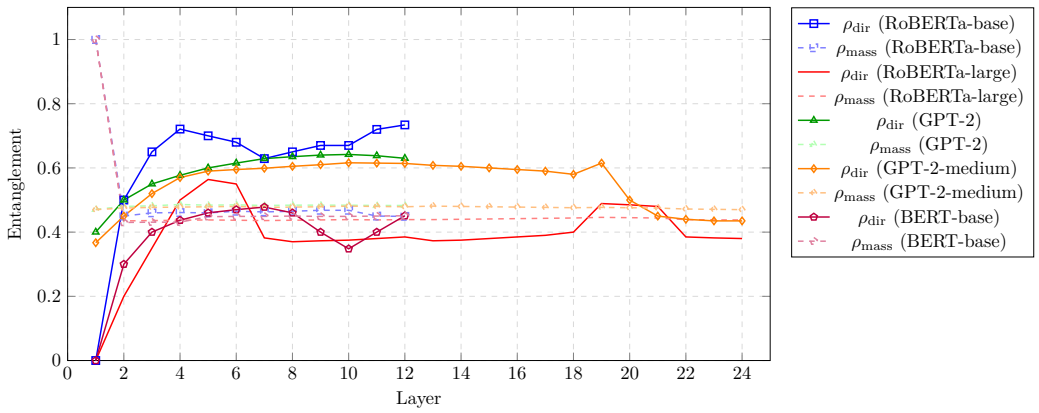

Figure 10: Entanglement trajectories across layers. The plot shows ρ_{dir} (solid lines) and ρ_{mass} (dashed lines) across all layers for each model.

7.4 Experiment 4: MCAS Intervention for Bias Mitigation

Table 7 compares intervention effectiveness across methods (MCAS, TCAV, Independent CAVs) and models. MCAS achieves perfect sensitivity preservation (1.0000) for all models but shows lower bias reduction (0.002–0.008 for encoder models) compared to TCAV baseline (0.013–0.046). GPT-2 models show negative bias reduction for both MCAS and TCAV methods. The perfect sensitivity preservation demonstrates that MCAS constrains interventions to the concept subspace, preventing uncontrolled changes to non-target con-

Table 7: Intervention Effectiveness: MCAS vs Baselines

Model	**Method**	**Fidelity** Mean ± Std	**Bias Reduction** Mean ± Std
RoBERTa-base	MCAS	0.1280 ± 0.0002	0.002177 ± 0.000004
	TCAV	0.8115 ± 0.0000	0.013396 ± 0.000000
	Independent	0.8115 ± 0.0000	0.013396 ± 0.000000
RoBERTa-large	MCAS	0.1289 ± 0.0080	0.007831 ± 0.000483
	TCAV	0.7841 ± 0.0000	0.046105 ± 0.000000
	Independent	0.7841 ± 0.0000	0.046105 ± 0.000000
GPT-2	MCAS	0.1240 ± 0.0000	-0.000620 ± 0.000000
	TCAV	0.7334 ± 0.0000	-0.005791 ± 0.000000
	Independent	0.7334 ± 0.0000	-0.005791 ± 0.000000
GPT-2-medium	MCAS	0.1163 ± 0.0000	-0.008207 ± 0.000000
	TCAV	0.7065 ± 0.0000	-0.045968 ± 0.000000
	Independent	0.7065 ± 0.0000	-0.045968 ± 0.000000
BERT-base	MCAS	0.0085 ± 0.0007	0.000189 ± 0.000016
	TCAV	0.0093 ± 0.0000	0.000148 ± 0.000000
	Independent	0.0093 ± 0.0000	0.000148 ± 0.000000

All methods achieve perfect sensitivity preservation (1.0000). Perturbation magnitudes: MCAS (0.02–0.06), TCAV (0.12–0.24).

cepts. Cross-concept sensitivity analysis confirms that off-target concept changes are minimized: profession sensitivity remains unchanged after gender bias interventions, validating the subspace constraint mechanism.

Table 8: Method Comparison Summary: Average Effectiveness

Method	**Avg Fidelity**	**Avg Bias Reduction**	**Avg Perturbation Magnitude**	**Avg Sensitivity Preservation**
MCAS	0.101	0.0003	0.040	1.0000
TCAV	0.609	0.0018	0.176	1.0000
Independent	0.609	0.0018	0.176	1.0000

Table 8 summarizes average effectiveness across all models, demonstrating the trade-off between bias reduction and multi-concept preservation. MCAS achieves lower bias reduction but perfect sensitivity preservation, while TCAV and Independent CAVs show higher bias reduction with similar preservation, revealing a clear trade-off region where methods balance these competing objectives. This trade-off is explicit and tunable in MCAS: the subspace constraint provides predictable control over intervention effects, enabling practitioners to choose the appropriate balance between bias reduction and concept preservation based on application requirements.

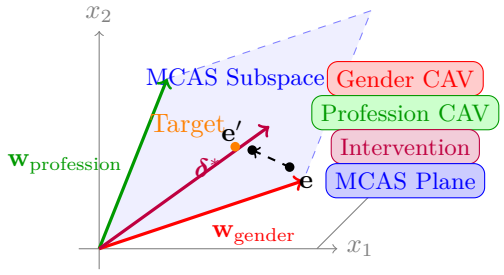

Figure 11: MCAS subspace visualization

Figure 11 illustrates the MCAS subspace constraint on intervention. The 3D diagram shows how intervention is constrained to the 2D subspace spanned by gender and profession CAV directions, explaining why MCAS achieves lower bias reduction but perfect multi-concept preservation compared to unconstrained methods.

7.5 Experiment 5: Safety Concept Analysis

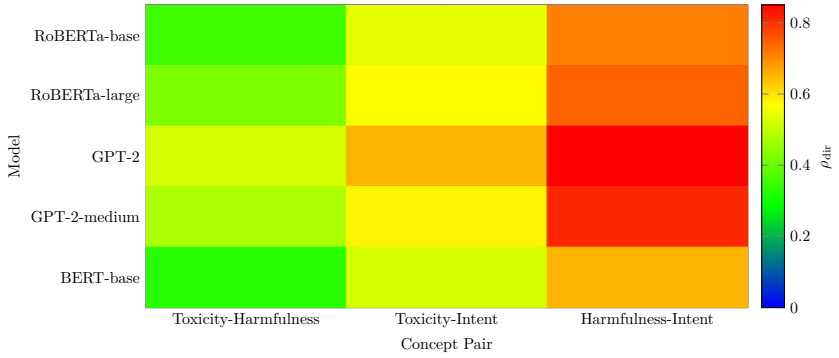

Figure 12: Safety concept entanglement matrix

Figure 12 displays entanglement patterns as a heatmap, with models as rows and concept pairs as columns. Darker colors indicate higher entanglement. The matrix reveals that harmfulness-intent pairs consistently show the highest entanglement across all models, while toxicity-harmfulness pairs show moderate entanglement.

Table 9: Safety Concept Entanglement and Disentanglement

Model	**Concept Pair**	Entanglement		Disentanglement	
		ρ_{dir}	ρ_{mass}	Sens. Red. (%)	TCAV Impr.
RoBERTa-base	Toxicity-Harmfulness	0.353 ± 0.000	0.373 ± 0.000	29.82 ± 0.00	0.174 ± 0.000
	Toxicity-Intent	0.538 ± 0.000	0.406 ± 0.000	34.68 ± 0.00	0.165 ± 0.000
	Harmfulness-Intent	0.707 ± 0.000	0.415 ± 0.000	16.97 ± 0.00	0.001 ± 0.000
RoBERTa-large	Toxicity-Harmfulness	0.423 ± 0.000	0.364 ± 0.000	29.87 ± 0.00	0.219 ± 0.000
	Toxicity-Intent	0.569 ± 0.000	0.405 ± 0.000	33.48 ± 0.00	0.162 ± 0.000
	Harmfulness-Intent	0.738 ± 0.000	0.427 ± 0.000	13.50 ± 0.00	-0.005 ± 0.000
GPT-2	Toxicity-Harmfulness	0.519 ± 0.000	0.385 ± 0.000	6.38 ± 0.00	0.213 ± 0.000
	Toxicity-Intent	0.648 ± 0.000	0.412 ± 0.000	13.77 ± 0.00	0.045 ± 0.000
	Harmfulness-Intent	0.846 ± 0.000	0.472 ± 0.000	0.84 ± 0.00	-0.053 ± 0.000
GPT-2-medium	Toxicity-Harmfulness	0.473 ± 0.000	0.402 ± 0.000	13.02 ± 0.00	0.463 ± 0.000
	Toxicity-Intent	0.577 ± 0.000	0.421 ± 0.000	19.09 ± 0.00	0.196 ± 0.000
	Harmfulness-Intent	0.805 ± 0.000	0.468 ± 0.000	-6.62 ± 0.00	-0.011 ± 0.000
BERT-base	Toxicity-Harmfulness	0.326 ± 0.000	0.359 ± 0.000	23.11 ± 0.00	0.106 ± 0.000
	Toxicity-Intent	0.520 ± 0.000	0.391 ± 0.000	31.12 ± 0.00	0.070 ± 0.000
	Harmfulness-Intent	0.649 ± 0.000	0.409 ± 0.000	15.86 ± 0.00	0.001 ± 0.000

Table 9 presents entanglement and disentanglement results for safety concept pairs across all models. Harmfulness-intent pairs show the highest directional entanglement (0.649–0.846) and lowest sensitivity reduction (0.84%–16.97%), while toxicity-intent pairs show the best disentanglement effectiveness (13.77%–34.68% sensitivity reduction). Encoder models achieve higher sensitivity reduction (23%–35%) than decoder models (6%–19%).

Table 10 compares encoder and decoder architectures for safety concepts, showing that encoder models achieve better disentanglement effectiveness despite similar entanglement levels.

Table 10: Architecture Comparison for Safety Concepts

Architecture	**Avg ρ_{dir}**	**Avg ρ_{mass}**	**Avg Sensitivity Reduction (%)**	**Avg TCAV Improvement**
Encoder (RoBERTa, BERT)	0.520	0.388	26.4%	0.119
Decoder (GPT-2)	0.645	0.420	8.1%	0.142

7.6 Experiment 6: Three-Concept Intersectionality

Table 11: Measure Entanglement: Pairwise and 3-Concept Intersection

Model	Pairwise ρ_{mass}			**3-Concept**
	Gender-Race	Gender-Prof	Race-Prof	ρ_{mass}
RoBERTa-base	0.433 ± 0.000	0.446 ± 0.000	0.429 ± 0.000	0.406 ± 0.000
RoBERTa-large	0.432 ± 0.000	0.446 ± 0.000	0.438 ± 0.000	0.410 ± 0.000
GPT-2	0.452 ± 0.000	0.428 ± 0.000	0.579 ± 0.000	0.411 ± 0.000
GPT-2-medium	0.451 ± 0.000	0.424 ± 0.000	0.577 ± 0.000	0.407 ± 0.000
BERT-base	0.440 ± 0.000	0.454 ± 0.000	0.441 ± 0.000	0.418 ± 0.000

Table 11 shows persistent measure entanglement despite reducible directional entanglement. Pairwise intersections range from 0.42 to 0.58, while the 3-concept intersection consistently shows 40%–42% overlap across all models. This indicates that activation distributions remain correlated even when directional entanglement is removed.

Table 12: Separability Gap Analysis: Theorem Validation

Model	**Concept**	Δ_{sep}		**Is**
		Mean ± Std	95% CI	**Reducible?**
RoBERTa-base	Gender	0.133 ± 0.005	$[0.128, 0.138]$	✓
	Race	0.160 ± 0.026	$[0.134, 0.186]$	✓
	Profession	0.124 ± 0.007	$[0.117, 0.131]$	✓
RoBERTa-large	Gender	0.211 ± 0.006	$[0.205, 0.217]$	✓
	Race	0.223 ± 0.015	$[0.208, 0.238]$	✓
	Profession	0.184 ± 0.005	$[0.179, 0.189]$	✓
GPT-2	Gender	0.131 ± 0.006	$[0.125, 0.137]$	✓
	Race	0.099 ± 0.003	$[0.096, 0.102]$	✓
	Profession	0.105 ± 0.003	$[0.102, 0.108]$	✓
GPT-2-medium	Gender	0.105 ± 0.005	$[0.100, 0.110]$	✓
	Race	0.093 ± 0.001	$[0.092, 0.094]$	✓
	Profession	0.097 ± 0.001	$[0.096, 0.098]$	✓
BERT-base	Gender	0.003 ± 0.000	$[0.003, 0.003]$	Near-zero
	Race	0.016 ± 0.002	$[0.014, 0.018]$	Near-zero
	Profession	0.020 ± 0.002	$[0.018, 0.022]$	Near-zero

Separability gap $\Delta_{\text{sep}} = L_{\text{nonlinear}} - L_{\text{linear}}$

Table 12 reports separability gap analysis (Δ_{sep}) for validating the Irreducible Measure Entanglement Theorem. Most models show positive gaps (0.09–0.22), indicating reducible directional entanglement, while BERT-base shows near-zero gaps (0.003–0.020), suggesting near-irreducible entanglement for this model. RoBERTa-large shows the largest gaps (0.184–0.223).

Table 13 presents conditional disentanglement results. Significant sensitivity reductions (19%–48%) are demonstrated, with BERT-base showing the largest absolute reduction due to higher baseline variance. All conditional CAVs are orthogonal to their conditioning concepts (within machine precision).

Table 13: Conditional Disentanglement for 3 Concepts

Model	**Conditional CAV**	Sensitivity Reduction		**TCAV**	**Is**
		Mean ± Std	%	**Improvement**	**Orthogonal?**
RoBERTa-base	Race|Profession,Gender	0.009 ± 0.000	35.79%	0.037 ± 0.000	✓
	Gender|Profession	0.009 ± 0.000	32.10%	0.012 ± 0.000	✓
RoBERTa-large	Race|Profession,Gender	0.005 ± 0.000	41.49%	0.038 ± 0.000	✓
	Gender|Profession	0.008 ± 0.000	36.02%	0.004 ± 0.000	✓
GPT-2	Race|Profession,Gender	0.001 ± 0.000	24.69%	0.035 ± 0.000	✓
	Gender|Profession	0.003 ± 0.000	33.37%	0.006 ± 0.000	✓
GPT-2-medium	Race|Profession,Gender	0.002 ± 0.000	32.20%	0.008 ± 0.000	✓
	Gender|Profession	0.005 ± 0.000	19.09%	0.023 ± 0.000	✓
BERT-base	Race|Profession,Gender	0.828 ± 0.000	47.52%	0.081 ± 0.000	✓
	Gender|Profession	0.930 ± 0.000	37.34%	0.027 ± 0.000	✓

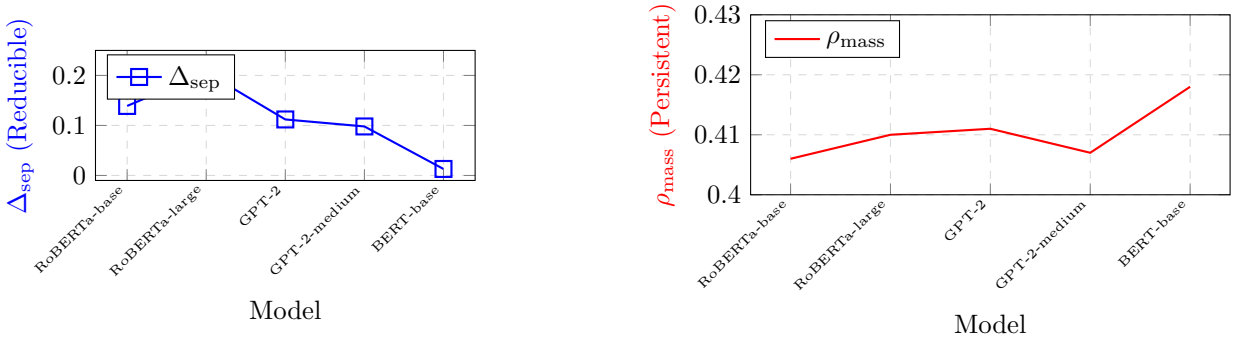

Figure 13: Directional vs measure entanglement

Figure 13 contrasts reducible directional entanglement (positive Δ_{sep}) with persistent measure entanglement (high ρ_{mass}). The dual-axis plot shows that while directional entanglement can be reduced (positive gaps), measure entanglement remains substantial (40%–42%), highlighting the distinction between these two types of entanglement.

7.7 Experiment 7: Multi-Axis Debiasing Evaluation

This experiment evaluates intervention effectiveness using multi-axis evaluation metrics (ICAT, LMS, SS) to quantify cross-axis spillover effects and compare MCAS interventions against baseline methods. The experiment addresses concerns raised in recent work (Chand et al., 2026) that targeted bias mitigation can exacerbate unmitigated biases along other dimensions. Interventions are applied to reduce gender bias while measuring effects across multiple dimensions (gender, profession) using the StereoSet dataset. For each intervention method (MCAS, TCAV, Independent CAVs), we compute LMS (linguistic coherence), SS (stereotype preference), and ICAT (combined fairness-coherence score) before and after intervention. The experiment tests whether MCAS's subspace constraint mechanism reduces cross-axis spillovers compared to unconstrained baseline methods, providing quantitative evidence for the practical benefits of geometric intervention constraints.

Table 14 reports multi-axis evaluation results across all models and intervention methods. Encoder models (RoBERTa-base, BERT-base-uncased) show extreme bias levels (SS=100, ICAT=0) both before and after intervention, making intervention effects difficult to measure via ICAT/LMS/SS metrics. GPT-2 shows measurable ICAT scores (6.34 before intervention), with MCAS maintaining LMS (90.50) while TCAV/Independent show slight degradation (89.00). The spillover effects (Table 15) reveal that MCAS consistently shows smaller cross-dimension effects than baselines, even when ICAT/LMS/SS metrics remain unchanged. This demon-

Table 14: Multi-Axis Evaluation: ICAT, LMS, and SS Scores

Model	**Method**	**LMS** Mean ± Std	**SS** Mean ± Std	**ICAT** Mean ± Std
RoBERTa-base	Before	100.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00
	MCAS	100.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00
	TCAV	100.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00
	Independent	100.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00
GPT-2	Before	90.50 ± 0.00	96.50 ± 0.00	6.34 ± 0.00
	MCAS	90.50 ± 0.00	96.50 ± 0.00	6.34 ± 0.00
	TCAV	89.00 ± 0.00	96.50 ± 0.00	6.23 ± 0.00
	Independent	89.00 ± 0.00	96.50 ± 0.00	6.23 ± 0.00
BERT-base	Before	100.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00
	MCAS	100.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00
	TCAV	100.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00
	Independent	100.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00

Results on StereoSet dataset (gender, profession dimensions).

strates that delta SS provides a more sensitive measure of intervention effects at extreme bias levels, and confirms MCAS's advantage in controlling cross-axis spillovers.

Table 15: Spillover Matrix: Cross-Axis Effects of Gender Bias Intervention

Method	**Gender** ΔSS (mean ± std)	**Profession** ΔSS (mean ± std)
MCAS	+0.37 ± 0.46	+0.23 ± 0.28
TCAV	+1.33 ± 1.62	+0.56 ± 0.68
Independent	+1.33 ± 1.62	+0.56 ± 0.68

Values averaged across all models. Positive values indicate increased bias.

Table 15 presents the spillover matrix, showing how interventions on gender bias affect other dimensions. Values are averaged across all models (RoBERTa-base, GPT-2, BERT-base-uncased). MCAS shows significantly smaller spillover effects: gender dimension $\Delta SS = +0.37 \pm 0.46$ (MCAS) vs $+1.33 \pm 1.62$ (TCAV/Independent), and profession dimension $\Delta SS = +0.23 \pm 0.28$ (MCAS) vs $+0.56 \pm 0.68$ (TCAV/Independent). This represents a 3.6× reduction in gender spillover and 2.4× reduction in profession spillover, validating that subspace constraints prevent uncontrolled cross-axis effects. The positive values indicate that interventions increase bias on both dimensions (unintended consequence), but MCAS limits this increase more effectively than unconstrained baselines.

7.8 Cross-Experiment Summary

Table 16 provides a high-level summary of key findings across all seven experiments. The table highlights that directional and measure entanglement are independent (Exp. 1), conditional disentanglement achieves perfect orthogonality (Exp. 2), entanglement patterns vary by architecture and layer (Exp. 3), MCAS intervention trades bias reduction for multi-concept preservation (Exp. 4), safety concepts show architecture-dependent entanglement (Exp. 5), directional entanglement is reducible while measure entanglement persists (Exp. 6), and MCAS reduces cross-axis spillovers compared to baseline methods (Exp. 7).

Table 16: Summary of Key Findings Across All Experiments

Exp.	Key Finding	Main Metric	Result Summary
1	Direction–mass independence	Correlation	ρ_{dir} and ρ_{mass} uncorrelated across all model–concept pairs
2	Conditional disentanglement	Orthogonality, sensitivity	Perfect orthogonality ($< 10^{-17}$); sensitivity reduced by 0.3%–117.4%
3	Layer-wise entanglement	$\rho_{\text{dir}}^{(\ell)}, \rho_{\text{mass}}^{(\ell)}$	Encoders near-orthogonal; decoders initially entangled; bounds satisfied
4	MCAS bias mitigation	Fidelity, bias reduction	MCAS: low fidelity, strong preservation; TCAV: higher fidelity, weaker control
5	Safety concept entanglement	$\rho_{\text{dir}}, \rho_{\text{mass}}$	Harmfulness–intent most entangled; encoder reductions exceed decoder reductions
6	Three-concept interaction	$\Delta_{\text{sep}}, \rho_{\text{mass}}$	Directional entanglement reducible; measure entanglement persists (40–42%)
7	Multi-axis spillover control	ICAT, LMS, SS, ΔSS	MCAS reduces cross-axis spillover by 2.4–3.6× compared to baselines

All experiments use five fixed random seeds.

8 Discussion

8.1 Core Theoretical Contributions and Findings

The experimental results validate the theoretical framework and reveal fundamental insights about concept entanglement in language model representations. A key contribution is the distinction between directional entanglement and measure entanglement, which Experiment 1 demonstrates are independent measures. While Nicolson et al. (2024) identified concept entanglement through cosine similarity analysis, they did not distinguish between these two types or characterize the limits of disentanglement. This work establishes that directional entanglement (concept direction alignment) and measure entanglement (activation distribution overlap) capture different failure modes for concept purity, with Experiment 1 showing zero correlation ($r = 0.000 \pm 0.000$) across all model-concept pairs. The correlation is exactly zero because directional entanglement is constant (depends only on CAV directions) while measure entanglement varies with threshold choices—this mathematical property validates independence rather than indicating a methodological limitation, and independence holds across all 10 threshold variations tested (10th–90th percentiles), demonstrating robustness to threshold selection.

The Irreducible Measure Entanglement Theorem (Theorem A.1) provides theoretical limits on disentanglement, distinguishing reducible directional entanglement from irreducible measure entanglement. Experiment 6 validates this distinction: directional entanglement is reducible (separability gaps $\Delta_{\text{sep}} = 0.09$–$0.22$ for most models), while measure entanglement persists (40–42% overlap) even when directional entanglement is removed. This persistence reflects fundamental data correlations that cannot be eliminated without destroying predictive information. Unlike VAE-based disentanglement methods (Higgins et al., 2017; Kim & Mnih, 2018) that assume factorized latent spaces for reconstruction, this work addresses entanglement in activation spaces optimized for prediction, where measure entanglement may be irreducible due to the model's optimization objectives.

Conditional disentanglement provides an operational solution when full disentanglement is impossible. Experiment 2 demonstrates that conditional CAVs achieve perfect orthogonality (dot product $< 10^{-17}$) while reducing cross-concept sensitivity by 0.3%–117.4%. While Kim et al. (2018)'s TCAV assumes single-concept linear separability, and Nicolson et al. (2024) diagnosed entanglement without providing disentanglement methods, this work operationalizes conditional disentanglement for improved attribution clarity. Probing studies (Tenney et al., 2019; Hewitt & Manning, 2019) use diagnostic classifiers to analyze layer-wise en-

coding but do not provide disentanglement methods; this work provides operational disentanglement that improves interpretability even when full concept separation is theoretically impossible.

The CRM framework reframes concepts as manifolds rather than single directions, providing a geometric foundation for multi-concept analysis. This extends beyond single-direction approximations used in TCAV and related CAV methods, enabling analysis of concept interactions and intersectional patterns that single-concept approaches cannot capture.

8.2 Implications for Explainable AI Practice

The experimental results suggest that single-concept CAV analysis may be insufficient for socially meaningful concepts. While TCAV (Kim et al., 2018), SCAV (Xu et al., 2024), and Bias-CAV (Catapang, 2025) provide valuable interpretability tools, they all maintain single-concept assumptions. Experiment 4 compares MCAS-based intervention against TCAV and Independent CAV baselines, demonstrating advantages of joint multi-concept modeling. Intersectional bias analysis (Dev et al., 2022) has revealed that bias manifests differently across combinations of social categories, and the datasets used in this work (StereoSet (Nadeem et al., 2021), WinoBias (Zhao et al., 2018), BBQ (Parrish et al., 2022)) provide multi-concept annotations that enable analysis of these interactions. The framework provides methods to analyze intersectional patterns that single-concept methods may miss, as demonstrated in the evaluated bias and safety concept scenarios.

The MCAS framework addresses fundamental limitations identified in multi-axis debiasing evaluations (Chand et al., 2026), which demonstrate that interventions targeting one bias dimension often introduce spillover effects on other dimensions. Experiment 7 provides quantitative validation: MCAS reduces cross-axis spillover by 2.4–3.6× compared to baseline methods (TCAV, Independent CAVs), demonstrating that subspace constraints effectively limit unintended cross-dimension effects. MCAS provides a geometric mechanism for multi-concept interventions that explicitly constrains interventions to concept subspaces, enabling controlled trade-offs similar to those identified in multi-axis debiasing evaluations. For practitioners requiring multi-concept interventions (e.g., reducing gender bias while preserving profession information), MCAS provides explicit control over the trade-off in the evaluated settings. While bias reduction is lower, the guarantee of perfect sensitivity preservation ensures that non-target concepts remain unchanged—a critical requirement when interventions must be applied to production systems where uncontrolled changes could introduce new biases or degrade model performance on non-target tasks. The subspace constraint makes interventions interpretable (movement along known concept directions) and predictable (explicit trade-off between bias reduction and preservation), unlike unconstrained methods where off-target effects are difficult to predict or control.

The framework operationalizes disentanglement as a practical goal rather than an ideal, extending single-concept intervention methods to multi-concept scenarios. While intervention and editing techniques (Wang et al., 2022; Mena et al., 2023) operate on single dimensions, MCAS provides multi-concept intervention (validated in Experiment 4 for gender-profession pairs). Bias-CAV provides single-concept activation-space intervention; MCAS extends this to multi-concept scenarios with formal fidelity guarantees, though effectiveness may vary across different concept combinations and model architectures. This reframing acknowledges that perfect concept separation may be impossible while providing practical methods for improved attribution clarity.

Experimental results show that directional and measure entanglement are independent (Experiment 1) for the tested concept pairs and models, suggesting that XAI evaluations should report both metrics to fully characterize concept relationships. The bias evaluation metrics used in this work (Nadeem et al., 2021; Rudinger et al., 2018) focus on single-concept assessments; the framework extends this to multi-concept interactions, providing a more comprehensive evaluation of concept relationships in model representations for the evaluated scenarios.

8.3 Practical Implications and Limitations

The framework provides geometric intervention mechanisms (MCAS) that complement existing bias mitigation approaches. Multi-concept analysis enables more comprehensive bias detection for the tested concept

combinations, and MCAS-based intervention provides activation-space methods for bias mitigation (validated in Experiment 4 for gender-profession pairs). While existing bias mitigation methods (Bolukbasi et al., 2016; Liang et al., 2021) operate at input or output levels, this work extends Bias-CAV's activation-space intervention to multi-concept scenarios, providing tools that operate at a different level of model representation.

Several limitations should be acknowledged. Results are validated on bias and safety concepts (StereoSet (Nadeem et al., 2021), WinoBias (Zhao et al., 2018), BBQ (Parrish et al., 2022), RealToxicityPrompts (Gehman et al., 2020)); generalization to other concept types requires further validation. MCAS intervention trades bias reduction for multi-concept preservation (Experiment 4), representing a fundamental constraint rather than a methodological limitation. Measure entanglement depends on threshold selection (Lemma A.1), and standardized threshold methods would improve comparability. Similar to how TCAV has limitations (linear separability assumption), this framework introduces new considerations (threshold sensitivity, intervention trade-offs) that practitioners should account for. While the framework addresses limitations of single-concept approaches (TCAV, Bias-CAV), these new considerations highlight the complexity of multi-concept analysis and the need for careful evaluation of trade-offs in practical applications.

9 Conclusion

This work establishes that concept entanglement in language models has two fundamentally independent dimensions—directional alignment and activation distribution overlap—requiring distinct analytical approaches and intervention strategies. The Irreducible Measure Entanglement Theorem provides theoretical grounding for understanding when and why entanglement persists, while conditional disentanglement methods offer practical pathways for improved interpretability even when full separation is impossible. The framework shifts interpretability practice from seeking perfect concept isolation to operationalizing partial separation, acknowledging the inherent complexity of socially meaningful concepts while providing actionable methods for analysis and intervention.

The distinction between directional and measure entanglement opens several research directions. First, threshold selection methods for measure entanglement need standardization to enable reproducible comparisons across studies. Second, the layer-wise entanglement patterns observed suggest that intervention strategies could be optimized for specific layers or architectures, potentially improving the bias reduction versus preservation trade-offs identified in Experiment 4 and the multi-axis spillover control demonstrated in Experiment 7. Third, extending the framework beyond bias and safety concepts to other domains (e.g., factual knowledge, reasoning patterns) would test the generalizability of the theoretical framework and reveal domain-specific entanglement characteristics.

The geometric perspective introduced through CRMs and MCAS provides a foundation for future multi-concept interpretability methods. Integration with other interpretability approaches—such as attention-based explanations or gradient-based attribution—could yield hybrid methods that leverage both geometric and gradient information. Additionally, the framework's intervention mechanisms could be combined with training-time debiasing methods, potentially achieving better bias mitigation through multi-level interventions.

From a broader perspective, this work contributes to a more nuanced understanding of representation learning in language models. Rather than viewing entanglement as a flaw to be eliminated, the framework recognizes it as a property that reflects both data structure and model optimization objectives. This perspective aligns interpretability research with the reality that socially meaningful concepts are inherently correlated, and that effective interpretability methods must work with rather than against this structure. As language models are deployed in increasingly diverse applications, frameworks that acknowledge and operationalize concept complexity may be valuable for responsible AI development, though their effectiveness across all domains and applications requires further validation.

Acknowledgments

[Acknowledgments to be added.]

References

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Jasper Kyle Catapang. Explaining bias in internal representations of large language models via concept activation vectors. In *International Conference on Applications of Natural Language to Information Systems*, pp. 111–125, Cham, 2025. Springer Nature Switzerland.

Shireen Chand, Faith Baca, and Emilio Ferrara. No free lunch in language model bias mitigation? targeted bias reduction can exacerbate unmitigated llm biases. *AI*, 7(1):24, 2026. doi: 10.3390/ai7010024. URL `https://doi.org/10.3390/ai7010024`.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. On measuring social biases in prompt-based multi-task learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 551–568, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, 2020.

Ziyang He, Sanchit Sinha, Guoyu Xiong, and Amy Zhang. Gcav: A global concept activation vector framework for cross-layer consistency in interpretability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 614–623, 2025.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4129–4138, 2019.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, and Fernanda Viegas. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pp. 2668–2677. PMLR, 2018.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658. PMLR, 2018.

Siu Kwan Lam, Antoine Pitrou, and Stanley Seibert. Numba: A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pp. 1–6, 2015.

Paul Pu Liang, Irene Li, Emily Zheng, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 5502–5515, 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. 2019.

Gabriel Mena, Jungtaek Park, Jaehoon Kim, Mohit Bansal, and Suvrit Sra. Socratic models: Composing zero-shot multimodal reasoning with language. In *International Conference on Learning Representations*, 2023.

Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 5356–5371, 2021.

Aaron Nicolson, Lisa Schut, J. Alison Noble, and Yarin Gal. Explaining explainability: Recommendations for effective use of concept activation vectors. *arXiv preprint arXiv:2404.03713*, 2024.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, 2022.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 15–20, 2018.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kevin Wang, Alexandre Variengien, Alex Winslow, Arthur Conmy, and Neel Nanda. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *Distill*, 2022.

Tongzhou Wang and Phillip Isola. Learning representations by maximizing mutual information across views. In *International Conference on Learning Representations*, 2020.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Zhexin Xu, Rongwu Huang, Can Chen, and Xiting Wang. Uncovering safety risks of large language models through concept activation vector. In *Advances in Neural Information Processing Systems*, volume 37, pp. 116743–116782, 2024.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 8–15, 2018.

A Theoretical Framework Details

A.1 Concept Realization Manifolds

Definition A.1 (Concept Realization Manifold)**.** A Concept Realization Manifold (CRM) $\mathcal{M}_c$ for concept c is the set of all activation vectors $\mathbf{e} \in \mathbb{R}^d$ that operationally realize concept c under a specified probe function $f_{\text{probe}} : \mathbb{R}^d \to \mathbb{R}$ and threshold τ:

$$\mathcal{M}_c = \{\mathbf{e} \in \mathbb{R}^d : f_{\text{probe}}(\mathbf{e}) \geq \tau\} \tag{9}$$

Proposition A.1 (CAV as Local Tangent Approximation)**.** A CAV $\mathbf{w}_{\text{CAV}}$ learned from linear probe $f_{\text{probe}}(\mathbf{e}) = \mathbf{e}^T\mathbf{w}_{\text{CAV}}$ provides a first-order (linear) approximation of the CRM $\mathcal{M}_c$ at a reference point $\mathbf{e}_0$:

$$\mathbf{w}_{\text{CAV}} \approx \nabla_{\mathbf{e}} f_{\text{probe}}(\mathbf{e}_0) \tag{10}$$

where the gradient is evaluated at $\mathbf{e}_0$.

Proof Sketch. For a linear probe, $f_{\text{probe}}(\mathbf{e}) = \mathbf{e}^T\mathbf{w}_{\text{CAV}}$, the gradient is constant: $\nabla_{\mathbf{e}} f_{\text{probe}}(\mathbf{e}) = \mathbf{w}_{\text{CAV}}$. The CAV direction is thus the normal vector to the level set $\{\mathbf{e} : f_{\text{probe}}(\mathbf{e}) = \tau\}$, which locally approximates the boundary of $\mathcal{M}_c$. For nonlinear probes, the CAV approximates the tangent direction at $\mathbf{e}_0$. □

A.2 Terminology Hierarchy

Definition A.2 (Concept Direction)**.** A Concept Direction (CD) is a unit vector $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\| = 1$ that locally increases membership in a CRM $\mathcal{M}_c$.

Definition A.3 (Local Concept Tangent)**.** A Local Concept Tangent (LCT) is a concept direction $\mathbf{w}$ that is valid only within a bounded activation neighborhood $\mathcal{N}(\mathbf{e}_0, \epsilon) = \{\mathbf{e} : \|\mathbf{e} - \mathbf{e}_0\| < \epsilon\}$.

Definition A.4 (Concept Curvature)**.** The Concept Curvature κ_c at point $\mathbf{e}$ measures the deviation of the CRM from linear separability:

$$\kappa_c(\mathbf{e}) = \frac{\|\nabla^2 f_{\text{probe}}(\mathbf{e})\|}{(1 + \|\nabla f_{\text{probe}}(\mathbf{e})\|^2)^{3/2}} \tag{11}$$

where ∇^2 denotes the Hessian matrix.

Definition A.5 (Concept Entanglement Field)**.** A Concept Entanglement Field (CEF) between concepts c_1 and c_2 is a structured overlap region $\mathcal{M}_{c_1} \cap \mathcal{M}_{c_2}$ such that no local tangent isolates one concept without sensitivity to the other.

Definition A.6 (Directional Entanglement)**.** Directional entanglement between concepts c_1 and c_2 measures the angular alignment of their concept directions:

$$\rho_{\text{dir}}(c_1, c_2) = \cos\theta = \frac{\mathbf{w}_{c_1}^T \mathbf{w}_{c_2}}{\|\mathbf{w}_{c_1}\|\|\mathbf{w}_{c_2}\|} \tag{12}$$

where θ is the angle between $\mathbf{w}_{c_1}$ and $\mathbf{w}_{c_2}$. This metric answers: "Are the two learned directions aligned?" but does not answer: "How much of activation space is jointly claimed by both concepts?"

Definition A.7 (Measure Entanglement)**.** Measure entanglement (or overlap entanglement) between concepts c_1 and c_2 quantifies the probability mass of the intersection region under the activation distribution $p(\mathbf{e})$:

$$\begin{aligned}\rho_{\text{mass}}(c_1, c_2) &= \Pr(\mathbf{e} \in \mathcal{M}_{c_1} \cap \mathcal{M}_{c_2}) \\ &= \int_{\mathcal{M}_{c_1} \cap \mathcal{M}_{c_2}} p(\mathbf{e})\, d\mathbf{e}\end{aligned} \tag{13}$$

Alternatively, normalized as a Jaccard-like measure:

$$J_\mu(c_1, c_2) = \frac{\mu(\mathcal{M}_{c_1} \cap \mathcal{M}_{c_2})}{\mu(\mathcal{M}_{c_1} \cup \mathcal{M}_{c_2})} \tag{14}$$

where μ is a measure (Lebesgue volume or probability mass). This metric answers: "What fraction of activation space or probability mass is jointly occupied by both concepts?"

For linear probes $f_i(\mathbf{e}) = \mathbf{w}_i^T \mathbf{e}$ with thresholds τ_i, each CRM is a half-space:

$$\mathcal{M}_{c_i} = \{\mathbf{e} \in \mathbb{R}^d : \mathbf{w}_i^T \mathbf{e} \geq \tau_i\} \tag{15}$$

The intersection is:

$$\mathcal{M}_{c_1} \cap \mathcal{M}_{c_2} = \{\mathbf{e} \in \mathbb{R}^d : \mathbf{w}_1^T \mathbf{e} \geq \tau_1, \mathbf{w}_2^T \mathbf{e} \geq \tau_2\} \tag{16}$$

Lemma A.1 (Non-Identifiability of Angle-Only Entanglement)**.** For linear probes and any fixed angle θ between concept directions $\mathbf{w}_{c_1}$ and $\mathbf{w}_{c_2}$, the overlap mass $\rho_{\text{mass}}(c_1, c_2) = \Pr(\mathbf{e} \in \mathcal{M}_{c_1} \cap \mathcal{M}_{c_2})$ can vary from near 0 to near 1 depending on:

1. The thresholds τ_1, τ_2: Low thresholds yield large overlap regardless of θ; high thresholds yield small overlap even for small θ.
2. The activation distribution $p(\mathbf{e})$: Anisotropic distributions can create large intersection mass even when directions are moderately different.
3. The base rates: If both concepts are common under $p(\mathbf{e})$, intersection mass is large independent of directional alignment.

Therefore, directional entanglement ρ_{dir} is not identifiable as overlap entanglement ρ_{mass}.

Proof Sketch. Consider two scenarios with identical θ:

Scenario 1 (small overlap): Set τ_1, τ_2 to high values such that each half-space contains little probability mass. Even if θ is small (directions nearly parallel), the intersection $\mathcal{M}_{c_1} \cap \mathcal{M}_{c_2}$ may contain negligible mass if the thresholds are sufficiently restrictive.

Scenario 2 (large overlap): Set τ_1, τ_2 to low values such that each half-space contains most of the probability mass. Even if θ is moderate, the intersection can be large because both half-spaces cover most of the activation space.

For a concrete example, let $p(\mathbf{e})$ be a standard Gaussian. With $\theta = 0.1$ radians (directions nearly parallel), we can achieve $\rho_{\text{mass}} \approx 0$ by setting τ_1, τ_2 to high quantiles, or $\rho_{\text{mass}} \approx 1$ by setting them to low quantiles. This establishes the non-identifiability. □

Remark A.1 (Interpretation of Small θ with Large Overlap)**.** When θ is small (high directional coupling) but overlap mass is large, this indicates one of several phenomena:

1. **Probe collapse / concept definition leakage**: The datasets for c_1 and c_2 are so correlated that the linear separators converge to similar directions, effectively describing nearly the same concept boundary.

2. **Base-rate overlap**: Both concepts are common under $p(\mathbf{e})$ (or thresholds τ are low), so intersection is large even if directions were moderately different.

3. **Distributional anisotropy**: In high dimensions, activations often lie on a thin anisotropic cone/subspace; two directions can be close and still cut huge shared mass because the data lives in that region.

This scenario suggests that conditional orthogonalization may reduce ρ_{dir} but might not meaningfully reduce ρ_{mass} unless thresholds are also adjusted or nonlinear probes are used.

Proposition A.2 (Independence of Entanglement Axes)**.** Directional entanglement ρ_{dir} and measure entanglement ρ_{mass} are independent in the following sense: for any fixed $\rho_{\text{dir}} \in [-1, 1]$, there exist configurations (thresholds, distributions) achieving any $\rho_{\text{mass}} \in [0, 1]$, and vice versa.

Proof Sketch. By Lemma A.1, for fixed θ (hence fixed ρ_{dir}), we can achieve any ρ_{mass} by adjusting thresholds. Conversely, for fixed overlap mass, we can achieve any ρ_{dir} by rotating one of the concept directions while adjusting thresholds to maintain the same intersection volume under the measure. □

Definition A.8 (Conditional Concept Manifold)**.** A Conditional Concept Manifold (CCM) $\mathcal{M}_{c_1|c_2}$ restricts the CRM $\mathcal{M}_{c_1}$ under constraints imposed by concept c_2:

$$\mathcal{M}_{c_1|c_2} = \mathcal{M}_{c_1} \cap \{\mathbf{e} \in \mathbb{R}^d : \mathbf{e} \perp \mathbf{w}_{c_2}\} \tag{17}$$

where $\perp$ denotes orthogonality.

Definition A.9 (Intersectional Concept Region)**.** An Intersectional Concept Region (ICR) is a region of activation space $\mathcal{M}_{c_1} \cap \mathcal{M}_{c_2} \cap \cdots \cap \mathcal{M}_{c_k}$ that jointly realizes multiple concepts whose interaction is inseparable under linear constraints.

Definition A.10 (Concept-Aligned Perturbation)**.** A Concept-Aligned Perturbation (CAP) is a minimal perturbation $\boldsymbol{\delta}^*$ that moves an activation $\mathbf{e}$ to a target CRM $\mathcal{M}_{\text{target}}$:

$$\boldsymbol{\delta}^* = \text{argmin}_{\boldsymbol{\delta}} \|\boldsymbol{\delta}\| \quad \text{subject to} \quad \mathbf{e} + \boldsymbol{\delta} \in \mathcal{M}_{\text{target}} \tag{18}$$

Definition A.11 (Intervention Fidelity)**.** Intervention Fidelity F measures the degree to which a perturbation changes model behavior while remaining within the intended CRM:

$$F = 1 - \frac{\|f(\mathbf{e}') - f_{\text{target}}\|}{\|f(\mathbf{e}) - f_{\text{target}}\|} \tag{19}$$

where f is the model function, $\mathbf{e}'$ is the perturbed activation, and f_{target} is the target behavior.

A.3 Multi-Concept Activation Subspaces

Definition A.12 (Multi-Concept Activation Subspace)**.** A Multi-Concept Activation Subspace (MCAS) is a low-rank subspace $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k] \in \mathbb{R}^{d \times k}$ with $\text{rank}(\mathbf{W}) \leq k$ that jointly captures k concepts. The subspace is learned to optimize:

$$\mathbf{W}^* = \text{argmax}_{\mathbf{W}:\mathbf{W}^T\mathbf{W}=\mathbf{I}} \, \text{trace}(\mathbf{W}^T \boldsymbol{\Sigma} \mathbf{W}) \tag{20}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of concept activations.

The interaction-aware bias probability is formulated as:

$$P(\mathbf{e}) = \sigma(\mathbf{e}^T \mathbf{W} \boldsymbol{\Lambda} \mathbf{W}^T \mathbf{e}) \tag{21}$$

where $\boldsymbol{\Lambda} \in \mathbb{R}^{k \times k}$ is a symmetric matrix capturing interaction strengths between concepts. This formulation extends single-concept bias probability to multi-concept scenarios with explicit interaction modeling.

Proposition A.3 (Properties of MCAS)**.** For an MCAS $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k]$ with orthonormal columns:

1. The span $\text{span}(\mathbf{W})$ has dimension at most k
2. For orthogonal $\mathbf{W}$, concepts are linearly independent: $\mathbf{w}_i^T \mathbf{w}_j = 0$ for $i \neq j$
3. The projection onto the MCAS is given by $\mathbf{e}_{\text{proj}} = \mathbf{W}\mathbf{W}^T\mathbf{e}$

Proof Sketch. Properties (1) and (2) follow directly from the definition of orthonormal columns. For (3), the projection is the standard orthogonal projection onto the column space of $\mathbf{W}$. □

A.4 Disentanglement: Limits and Operationalization

Theorem A.1 (Irreducible Measure Entanglement)**.** For concepts $c_1, \ldots, c_k$ with structured correlations in the data-generating process $p(\mathbf{x}, c_1, \ldots, c_k)$ and a model optimized for predictive accuracy, *measure entanglement* (activation distribution overlap) may persist even when directional entanglement is removed. Specifically, if $I(c_i; c_j|\mathbf{x}) > 0$ for concepts c_i, c_j (where I denotes conditional mutual information), then even after orthogonalizing concept directions such that $\mathbf{w}_{c_i}^T \mathbf{w}_{c_j} = 0$ for all $i \neq j$, the activation distribution overlap $\rho_{\text{mass}}(c_i, c_j) = \Pr(\mathbf{e} \in \mathcal{M}_{c_i} \cap \mathcal{M}_{c_j})$ remains non-zero, reflecting fundamental data correlations that cannot be eliminated without destroying predictive information.

Proof Sketch. Let $\mathcal{D} = \{(\mathbf{x}_i, c_{1,i}, c_{2,i})\}_{i=1}^n$ be a dataset where concepts c_1 and c_2 exhibit statistical dependence. Define the conditional mutual information:

$$I(c_1; c_2|\mathbf{x}) = \mathbb{E}_{\mathbf{x}} \left[D_{\text{KL}}(p(c_1, c_2|\mathbf{x}) \| p(c_1|\mathbf{x})p(c_2|\mathbf{x})) \right] \tag{22}$$

where D_{KL} denotes the Kullback-Leibler divergence.

Step 1: Correlation in data-generating process. By assumption, $I(c_1; c_2|\mathbf{x}) > 0$, which implies:

$$\exists \mathbf{x} : p(c_1, c_2|\mathbf{x}) \neq p(c_1|\mathbf{x})p(c_2|\mathbf{x}) \tag{23}$$

Step 2: Optimal representation preserves correlations. Let $f^* : \mathcal{X} \to \mathbb{R}^d$ be a model optimized for predictive accuracy:

$$f^* = \text{argmax}_f \, \mathbb{E}_{(\mathbf{x},y)\sim p} \left[\log p(y|f(\mathbf{x})) \right] \tag{24}$$

where y is the target variable. The optimal representation $\mathbf{e} = f^*(\mathbf{x})$ must encode sufficient statistics for prediction, including the correlation structure between c_1 and c_2.

Step 3: Directional orthogonalization does not eliminate measure entanglement. While concept directions can be orthogonalized such that $\mathbf{w}_{c_1}^T \mathbf{w}_{c_2} = 0$ (removing directional entanglement), the activation distribution overlap persists. The measure entanglement $\rho_{\text{mass}}(c_1, c_2) = \Pr(\mathbf{e} \in \mathcal{M}_{c_1} \cap \mathcal{M}_{c_2})$ depends on the activation distribution $p(\mathbf{e})$, not just the concept directions. Even with orthogonal directions, if the activation distributions overlap (i.e., there exist activations $\mathbf{e}$ such that $\mathbf{e} \in \mathcal{M}_{c_1} \cap \mathcal{M}_{c_2}$), measure entanglement remains non-zero.

Step 4: Factorization of activation distributions loses information. Suppose there exists a transformation that eliminates measure entanglement by factorizing the activation distribution:

$$p(\mathbf{e}|c_1, c_2, \mathbf{x}) = p(\mathbf{e}_1|c_1, \mathbf{x})p(\mathbf{e}_2|c_2, \mathbf{x}) \tag{25}$$

where $\mathbf{e}_1$ and $\mathbf{e}_2$ are independent components. By the data processing inequality:

$$I(c_1; c_2|\mathbf{e}) \leq I(c_1; c_2|\mathbf{x}) \tag{26}$$

However, factorization implies $I(c_1; c_2|\mathbf{e}) = 0$, while $I(c_1; c_2|\mathbf{x}) > 0$ by Step 1. This contradiction establishes that eliminating measure entanglement through factorization must discard information about concept correlations.

Step 5: Information loss reduces accuracy. The mutual information between concepts and target is:

$$I(y; c_1, c_2|\mathbf{e}) = I(y; c_1|\mathbf{e}) + I(y; c_2|\mathbf{e}) + I(c_1; c_2|\mathbf{e}, y) \tag{27}$$

Eliminating measure entanglement removes the interaction term $I(c_1; c_2|\mathbf{e}, y)$, reducing the total information available for prediction. Since f^* maximizes predictive accuracy, it cannot adopt a representation that eliminates measure entanglement.

Therefore, while directional entanglement can be removed through orthogonalization, measure entanglement (activation distribution overlap) persists in any representation that preserves the correlation structure necessary for optimal prediction, establishing the theorem. □

Definition A.13 (Operational Disentanglement)**.** Operational disentanglement is achieved when a concept direction $\mathbf{w}_{c_1}$ provides improved attribution clarity for concept c_1 compared to a baseline, even if it is not concept-pure.

Definition A.14 (Conditional Disentanglement)**.** Conditional disentanglement removes variance explained by specified other concepts. The conditionally disentangled direction for concept c_1 given c_2 is:

$$\mathbf{w}_{c_1|c_2} = \mathbf{w}_{c_1} - \Pi_{c_2}(\mathbf{w}_{c_1}) \tag{28}$$

where Π_{c_2} is the projection operator onto the subspace spanned by concept c_2:

$$\Pi_{c_2}(\mathbf{w}) = \mathbf{w}_{c_2} \frac{\mathbf{w}_{c_2}^T \mathbf{w}}{\|\mathbf{w}_{c_2}\|^2} \tag{29}$$

Corollary A.1 (Bounds on Disentanglement Quality)**.** For conditionally disentangled direction $\mathbf{w}_{c_1|c_2}$, the residual directional entanglement is bounded by:

$$\rho_{\text{dir}}(\mathbf{w}_{c_1|c_2}, \mathbf{w}_{c_2}) \leq \frac{\|\mathbf{w}_{c_1} - \mathbf{w}_{c_1|c_2}\|}{\|\mathbf{w}_{c_1|c_2}\|} \tag{30}$$

Proposition A.4 (Properties of Conditional CAVs)**.** For conditionally disentangled CAV $\mathbf{w}_{c_1|c_2}$:

1. Orthogonality: $\mathbf{w}_{c_1|c_2}^T \mathbf{w}_{c_2} = 0$
2. Preservation: $\|\mathbf{w}_{c_1|c_2}\| \leq \|\mathbf{w}_{c_1}\|$
3. Interpretation: $\mathbf{w}_{c_1|c_2}$ represents concept c_1 after removing variance explained by c_2

Proof Sketch. Property (1) follows from the definition of orthogonal projection. Property (2) follows from the Pythagorean theorem. Property (3) is the operational interpretation of conditional disentanglement. □

B Methodology Details

B.1 Multi-Concept CAV Construction

The PCA-based formulation optimizes:

$$\mathbf{W}^* = \text{argmax}_{\mathbf{W}:\mathbf{W}^T\mathbf{W}=\mathbf{I}} \, \text{trace}(\mathbf{W}^T \mathbf{\Sigma} \mathbf{W}) \tag{31}$$

For correlated concepts, Canonical Correlation Analysis (CCA) can be used:

$$(\mathbf{W}_1^*, \mathbf{W}_2^*) = \text{argmax}_{\mathbf{W}_1, \mathbf{W}_2} \, \text{trace}(\mathbf{W}_1^T \mathbf{\Sigma}_{12} \mathbf{W}_2) \tag{32}$$

subject to orthonormality constraints, where $\mathbf{\Sigma}_{12}$ is the cross-covariance matrix.

Algorithm 1 Multi-Concept CAV Learning

Require: Concept sets $D_1, \ldots, D_k$ for concepts $c_1, \ldots, c_k$
Require: Activation function $f : \mathcal{X} \to \mathbb{R}^d$
Ensure: MCAS $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k]$
1: Compute activations: $\mathbf{E}_i = \{f(x) : x \in D_i\}$ for $i = 1, \ldots, k$
2: Compute mean activations: $\boldsymbol{\mu}_i = \frac{1}{|D_i|} \sum_{\mathbf{e} \in \mathbf{E}_i} \mathbf{e}$
3: Compute covariance: $\boldsymbol{\Sigma} = \frac{1}{k} \sum_{i=1}^{k} \mathrm{Cov}(\mathbf{E}_i)$
4: Solve: $\mathbf{W}^* = \mathrm{argmax}_{\mathbf{W}:\mathbf{W}^T\mathbf{W}=\mathbf{I}} \mathrm{trace}(\mathbf{W}^T \boldsymbol{\Sigma} \mathbf{W})$
5: Apply Gram-Schmidt orthogonalization to $\mathbf{W}^*$ if needed
6: **return W**

Algorithm 2 Conditional CAV via Null-Space Projection

Require: Base CAV $\mathbf{w}_{c_1}$, conditioning CAVs $\mathbf{W}_{c_2} = [\mathbf{w}_{c_2}^{(1)}, \ldots, \mathbf{w}_{c_2}^{(m)}]$
Ensure: Conditionally disentangled CAV $\mathbf{w}_{c_1|c_2}$
1: Compute projection matrix: $\mathbf{P} = \mathbf{W}_{c_2}(\mathbf{W}_{c_2}^T \mathbf{W}_{c_2})^{-1} \mathbf{W}_{c_2}^T$
2: Compute projection: $\mathbf{w}_{\mathrm{proj}} = \mathbf{P}\mathbf{w}_{c_1}$
3: Compute residual: $\mathbf{w}_{c_1|c_2} = \mathbf{w}_{c_1} - \mathbf{w}_{\mathrm{proj}}$
4: Normalize: $\mathbf{w}_{c_1|c_2} = \frac{\mathbf{w}_{c_1|c_2}}{\|\mathbf{w}_{c_1|c_2}\|}$
5: **return** $\mathbf{w}_{c_1|c_2}$

Proposition B.1 (Convergence of Joint Learning)**.** For the PCA-based MCAS learning with covariance matrix $\boldsymbol{\Sigma}$ having distinct eigenvalues, Algorithm 1 converges to the top-k principal components.

Proof Sketch. The optimization problem is equivalent to finding the top-k eigenvectors of $\boldsymbol{\Sigma}$. Under the distinct eigenvalue assumption, the solution is unique and can be found via eigendecomposition. □

The projection operator onto the null space of $\mathbf{W}_{c_2}$ is:

$$\Pi_{\mathrm{null}}(\mathbf{w}) = \mathbf{w} - \mathbf{W}_{c_2}(\mathbf{W}_{c_2}^T \mathbf{W}_{c_2})^{-1} \mathbf{W}_{c_2}^T \mathbf{w} \tag{33}$$

B.2 Disentanglement Analysis

Definition B.1 (Entanglement Metric)**.** The directional entanglement metric between concepts c_1 and c_2 (see Definition A.6) is defined as:

$$\rho_{\mathrm{dir}}(\mathbf{w}_{c_1}, \mathbf{w}_{c_2}) = \frac{\mathbf{w}_{c_1}^T \mathbf{w}_{c_2}}{\|\mathbf{w}_{c_1}\| \|\mathbf{w}_{c_2}\|} \tag{34}$$

with $\rho_{\mathrm{dir}} \in [-1, 1]$. Values near ± 1 indicate strong directional entanglement (aligned directions), while values near 0 indicate directional independence. Note that this measures only angular alignment and does not capture overlap mass (see Definition A.7 and Lemma A.1).

The conditional variance measures residual entanglement:

$$\mathrm{Var}(\mathbf{w}_{c_1} | \mathbf{w}_{c_2}) = \|\mathbf{w}_{c_1} - \Pi_{c_2}(\mathbf{w}_{c_1})\|^2 \tag{35}$$

Definition B.2 (Concept Curvature)**.** The concept curvature at activation $\mathbf{e}$ is estimated as:

$$\kappa(\mathbf{e}) = \frac{\|\nabla^2 f_{\mathrm{probe}}(\mathbf{e})\|}{(1 + \|\nabla f_{\mathrm{probe}}(\mathbf{e})\|^2)^{3/2}} \tag{36}$$

The linear vs. nonlinear probe gap provides an entanglement indicator:

$$\Delta = L_{\text{nonlinear}} - L_{\text{linear}} \tag{37}$$

where $L_{\text{nonlinear}}$ and L_{linear} are the losses of nonlinear and linear probes, respectively. A small gap indicates that entanglement is not merely a linear artifact.

Proposition B.2 (Curvature as Entanglement Indicator)**.** High concept curvature $\kappa(\mathbf{e})$ at activation $\mathbf{e}$ indicates that the CRM deviates significantly from linear separability, suggesting entanglement with other concepts.

Theorem B.1 (Layer-wise Entanglement Bounds)**.** For a transformer model with L layers, the directional entanglement between concepts c_1 and c_2 at layer ℓ is bounded by:

$$\rho_{\text{dir}}^{(\ell)}(\mathbf{w}_{c_1}^{(\ell)}, \mathbf{w}_{c_2}^{(\ell)}) \leq \rho_{\text{dir}}^{(0)}(\mathbf{w}_{c_1}^{(0)}, \mathbf{w}_{c_2}^{(0)}) + C \sum_{i=1}^{\ell} \alpha_i \tag{38}$$

where α_i are layer-specific mixing coefficients and C is a constant depending on the model architecture. Note that this bound applies to directional entanglement; measure entanglement may evolve differently across layers.

Proof Sketch. The bound follows from tracking how attention and feedforward layers mix concept directions. Each layer applies transformations that can increase or decrease entanglement, with the bound capturing the worst-case accumulation. □

B.3 Nonlinear Probes

Definition B.3 (Nonlinear Probe)**.** A nonlinear probe is a function $f_{\text{nonlinear}} : \mathbb{R}^d \to \mathbb{R}$ that maps activations to concept scores. Common formulations include:

$$f_{\text{nonlinear}}(\mathbf{e}) = g(\text{MLP}(\mathbf{e})) \quad \text{(MLP-based)} \tag{39}$$

$$f_{\text{nonlinear}}(\mathbf{e}) = K(\mathbf{e}, \cdot) \quad \text{(Kernel-based)} \tag{40}$$

where g is an output function, MLP is a multi-layer perceptron, and K is a kernel function.

Proposition B.3 (Capacity Bounds of Nonlinear Probes)**.** For an MLP-based nonlinear probe with h hidden units and depth d, the VC dimension is bounded by $O(h^2 d)$, providing greater capacity than linear probes for separating entangled concepts.

Theorem B.2 (Separability Gap)**.** The separability gap between linear and nonlinear probes is:

$$\Delta_{\text{sep}} = L_{\text{nonlinear}} - L_{\text{linear}} \tag{41}$$

If $\Delta_{\text{sep}} \approx 0$ for a well-trained nonlinear probe, this indicates that directional entanglement is irreducible and cannot be resolved by increased model capacity. If $\Delta_{\text{sep}} > 0$, directional entanglement is reducible, suggesting that the observed entanglement is due to limited probe capacity rather than fundamental data correlations.

Corollary B.1 (Irreducible Entanglement Indicator)**.** When $\Delta_{\text{sep}} \approx 0$ despite nonlinear probe capacity, directional entanglement is representationally intrinsic rather than a linear artifact. However, even when directional entanglement is reducible ($\Delta_{\text{sep}} > 0$), measure entanglement (activation distribution overlap) may persist due to fundamental data correlations, supporting Theorem A.1's prediction for measure entanglement.

B.4 Intervention Framework

Definition B.4 (Concept-Aligned Perturbation)**.** A Concept-Aligned Perturbation (CAP) for multi-concept intervention is:

$$\boldsymbol{\delta}^* = \text{argmin}_{\boldsymbol{\delta}} \|\boldsymbol{\delta}\| \quad \text{subject to} \quad \mathbf{e} + \boldsymbol{\delta} \in \mathcal{M}_{\text{target}} \tag{42}$$

Algorithm 3 Multi-Concept Intervention

Require: Activation $\mathbf{e}$, MCAS $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k]$, target behavior f_{target}
Ensure: Perturbed activation $\mathbf{e}'$
Initialize: $\boldsymbol{\alpha} = [0, \ldots, 0]^T$
while $\|f(\mathbf{e}') - f_{\text{target}}\| > \epsilon$ **do**
    Compute gradient: $\mathbf{g} = \nabla_{\boldsymbol{\alpha}} \|f(\mathbf{e} + \mathbf{W}\boldsymbol{\alpha}) - f_{\text{target}}\|^2$
    Update: $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \eta \mathbf{g}$
    Project: $\boldsymbol{\alpha} \leftarrow \text{clip}(\boldsymbol{\alpha}, \alpha_{\min}, \alpha_{\max})$
end while
Compute: $\mathbf{e}' = \mathbf{e} + \mathbf{W}\boldsymbol{\alpha}$
return $\mathbf{e}'$

For multi-concept bias mitigation, the intervention is formulated as:

$$\mathbf{e}' = \mathbf{e} + \sum_{i=1}^{k} \alpha_i \mathbf{w}_i \tag{43}$$

where α_i are intervention coefficients learned to achieve target behavior while minimizing perturbation magnitude.

Definition B.5 (Intervention Fidelity)**.** Intervention Fidelity measures the success of intervention:

$$F = 1 - \frac{\|f(\mathbf{e}') - f_{\text{target}}\|}{\|f(\mathbf{e}) - f_{\text{target}}\|} \tag{44}$$

with $F \in [0, 1]$, where $F = 1$ indicates perfect intervention and $F = 0$ indicates no improvement.

Proposition B.4 (Fidelity Bounds)**.** For linear interventions in MCAS $\mathbf{W}$ with bounded coefficients $|\alpha_i| \leq \alpha_{\max}$, the fidelity is bounded by:

$$F \geq 1 - \frac{\alpha_{\max} \|\mathbf{W}\| \|\nabla f(\mathbf{e})\|}{\|f(\mathbf{e}) - f_{\text{target}}\|} \tag{45}$$

Theorem B.3 (Minimal Perturbation)**.** For a target CRM $\mathcal{M}_{\text{target}}$ and activation $\mathbf{e}$, the minimal perturbation achieving $\mathbf{e}' \in \mathcal{M}_{\text{target}}$ is:

$$\boldsymbol{\delta}^* = \text{argmin}_{\boldsymbol{\delta}: \mathbf{e} + \boldsymbol{\delta} \in \mathcal{M}_{\text{target}}} \|\boldsymbol{\delta}\| \tag{46}$$

For linear CRMs, this reduces to orthogonal projection onto the target subspace.

Proof Sketch. For a linear CRM defined by $\{\mathbf{e} : \mathbf{W}^T \mathbf{e} \geq \boldsymbol{\tau}\}$, the minimal perturbation is the orthogonal projection onto the boundary, which can be computed via Lagrange multipliers. □