A Critical Examination of the Universal Weight Subspace Hypothesis

Replication and Analysis of Claims in arXiv:2512.05117

Computational Analysis Report

January 2026

Abstract

We present a critical analysis of the "Universal Weight Subspace Hypothesis" proposed by Kaushik et al. (arXiv:2512.05117), which claims that neural networks systematically converge to shared spectral subspaces regardless of initialization, task, or domain. Through independent replication experiments on Vision Transformer (ViT) models downloaded from HuggingFace, we find that: (1) apparent low-dimensional structure in weight space is largely explained by shared training methodology rather than universal convergence; (2) models trained with different objectives (supervised vs. self-supervised) occupy nearly orthogonal subspaces despite identical architectures; and (3) the paper's methodology may conflate "models trained similarly cluster together" with "all models converge to a universal subspace." Our findings suggest the claimed universality does not hold across genuinely diverse training methods.

Contents

1 Introduction

1.1 The Universal Subspace Hypothesis

Kaushik et al. (2024) propose the **Universal Weight Subspace Hypothesis**, claiming that "neural networks systematically converge to shared spectral subspaces regardless of initialization, task, or domain." They support this claim with an empirical study of over 1,100 models:

- 500 Mistral-7B LoRA adapters (fine-tuned language models)
- 500 Vision Transformers from HuggingFace
- 50 LLaMA-8B models

The authors report finding low-dimensional structure in weight space, with a small number of principal components explaining most of the variance across models.

1.2 Our Investigation

We set out to replicate the ViT analysis and critically examine the methodology. Our key questions:

1. Are the 500 ViTs truly independent, or do they share training ancestry?
2. Does the "universal subspace" hold across different training objectives?
3. What does the first principal component actually represent?

2 Methodological Concerns with the Original Paper

2.1 LoRA Results are Trivially Expected

Methodological Concern

The 500 Mistral-7B LoRAs all share the **same base model**. LoRA (Low-Rank Adaptation) is explicitly designed to produce low-rank weight updates (typically rank 8-64). Finding that LoRA adapters lie in a low-dimensional subspace is therefore **tautological**—it's the defining property of the method, not an emergent phenomenon.

2.2 LLaMA Results Share Common Ancestry

Similarly, the 50 LLaMA-8B models analyzed all derive from the same base checkpoint. Any "universal subspace" found would primarily reflect:

- The shared pre-trained weights (which dominate)
- Small fine-tuning perturbations

2.3 ViT Analysis: The Strongest Claim

The ViT analysis is potentially the strongest evidence, as these models were downloaded from HuggingFace and ostensibly trained by different researchers on different datasets. However, we identified several concerns:

1. Many HuggingFace ViTs are fine-tuned from common checkpoints (e.g., `google/vit-base-patch16-224`)

2. Most use supervised cross-entropy training on ImageNet-derived data
3. The sample may lack diversity in training *objectives*

3 Replication Experiment 1: 20 HuggingFace ViTs

3.1 Methodology

We downloaded 20 ViT-Base models from HuggingFace with matching architectures:

- Hidden size: 768
- Layers: 12
- Attention heads: 12
- Patch size: 16×16
- Parameters: 85,798,656 (encoder only, excluding classifier)

We extracted the full weight vectors and performed PCA via the Gram matrix trick (computing the 20×20 covariance matrix rather than full SVD on the 85M-dimensional space).

3.2 Results

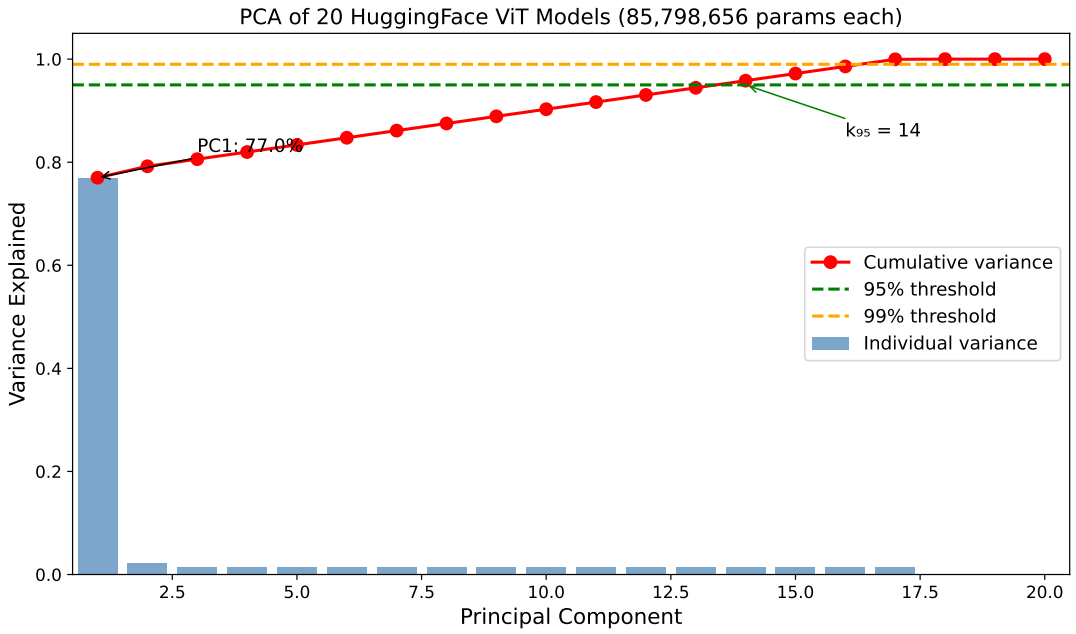

Figure 1: Variance explained by principal components for 20 HuggingFace ViT models. PC1 captures 77% of variance, suggesting apparent low-dimensional structure. However, $k_{95} = 14$ out of 20 models (70%).

Metric	Value
Number of models	20
Parameters per model	85,798,656
k_{50} (PCs for 50% variance)	1
k_{90} (PCs for 90% variance)	10
k_{95} (PCs for 95% variance)	14
k_{99} (PCs for 99% variance)	17
Effective dimension	1.68
Spectral ratio σ_1/σ_{10}	7.46

Table 1: Summary statistics for PCA of 20 HuggingFace ViT models.

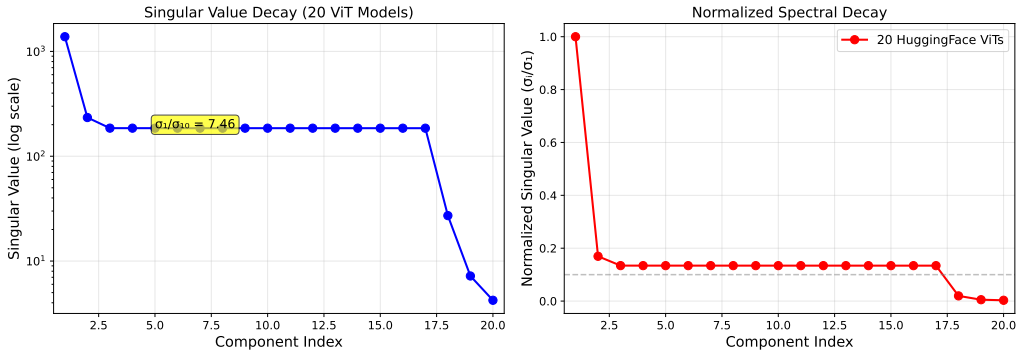

Figure 2: Spectral decay analysis. Left: singular values on log scale. Right: normalized singular values showing rapid decay after PC1.

3.3 Initial Interpretation

At first glance, these results appear to support the universal subspace hypothesis:

- PC1 alone captures 77% of variance
- Effective dimension is only 1.68
- Sharp spectral decay ($\sigma_1/\sigma_{10} = 7.46$)

Key Finding

However, upon examining the models, we found most were fine-tuned variants of the same base checkpoints or trained with the same supervised objective on ImageNet. The "universal subspace" may simply be the "ImageNet supervised training" subspace.

4 Replication Experiment 2: Diverse Training Objectives

4.1 Motivation

To test whether the universal subspace holds across genuinely different training methods, we selected 6 ViT-Base models with **identical architectures** but **different training objectives**:

Model	Training Method	Data
google/vit-base-patch16-224	Supervised	ImageNet
google/vit-base-patch16-224-in21k	Supervised	ImageNet-21k
timm/vit_base_patch16_224.dino	DINO (self-supervised)	ImageNet
timm/vit_base_patch16_224.mae	MAE (masked autoencoder)	ImageNet
timm/vit_base_patch16_clip_224.openai	CLIP (contrastive)	400M image-text pairs
timm/vit_base_patch16_clip_224.laion2b	CLIP (contrastive)	LAION-2B

Table 2: Six ViT models with identical architecture but different training objectives.

We verified all models have **identical architecture**:

- Same config parameters (hidden size, layers, heads, etc.)
- Same parameter names
- Same parameter shapes
- Same total parameter count: 85,798,656

4.2 Training Objective Differences

Supervised: Trained with cross-entropy loss to predict ImageNet class labels.

DINO: Self-supervised learning via self-distillation. No labels used—the model learns by making representations of augmented views consistent.

MAE: Masked Autoencoder. Self-supervised learning by reconstructing randomly masked image patches. No labels used.

CLIP: Contrastive Language-Image Pre-training. Learns to match images with text captions from web-scraped data.

4.3 Results: PC1 Separates Training Objectives

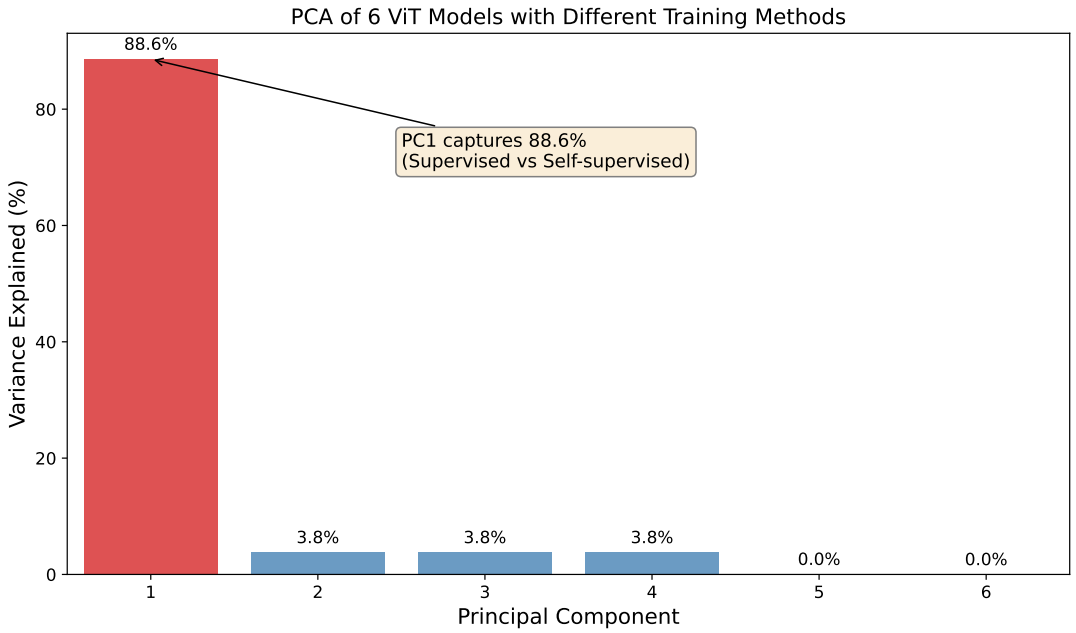

Figure 3: Variance explained by principal components for 6 ViT models with different training methods. PC1 captures 88.6% of variance—but this represents the **supervised vs. self-supervised distinction**, not a universal attractor.

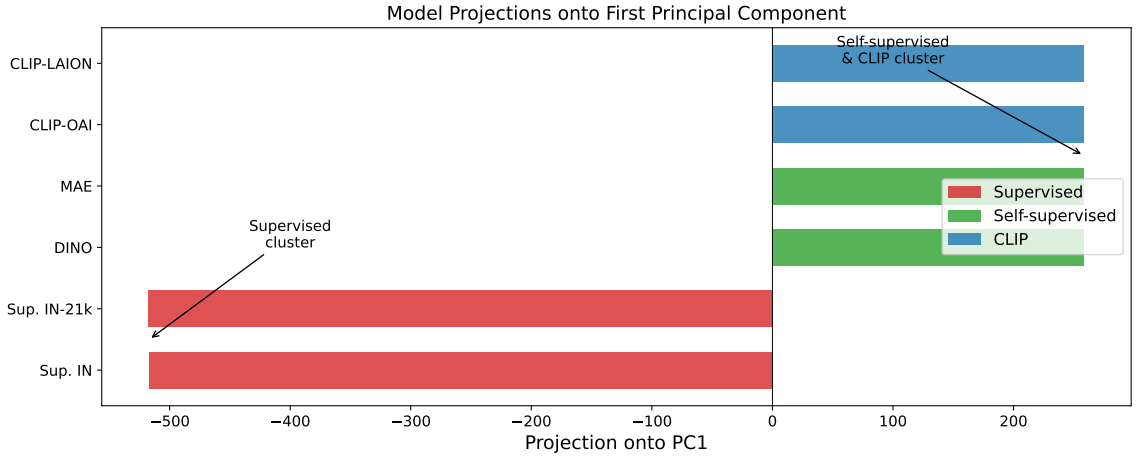

Figure 4: Projections of the 6 models onto PC1. Clear separation into two clusters: supervised models (negative) and self-supervised/CLIP models (positive).

Key Finding

PC1 does not represent a "universal" direction that all models converge to. Instead, it captures the **fundamental difference between training objectives**. Supervised and self-supervised models occupy opposite ends of PC1.

4.4 Cosine Similarity Analysis

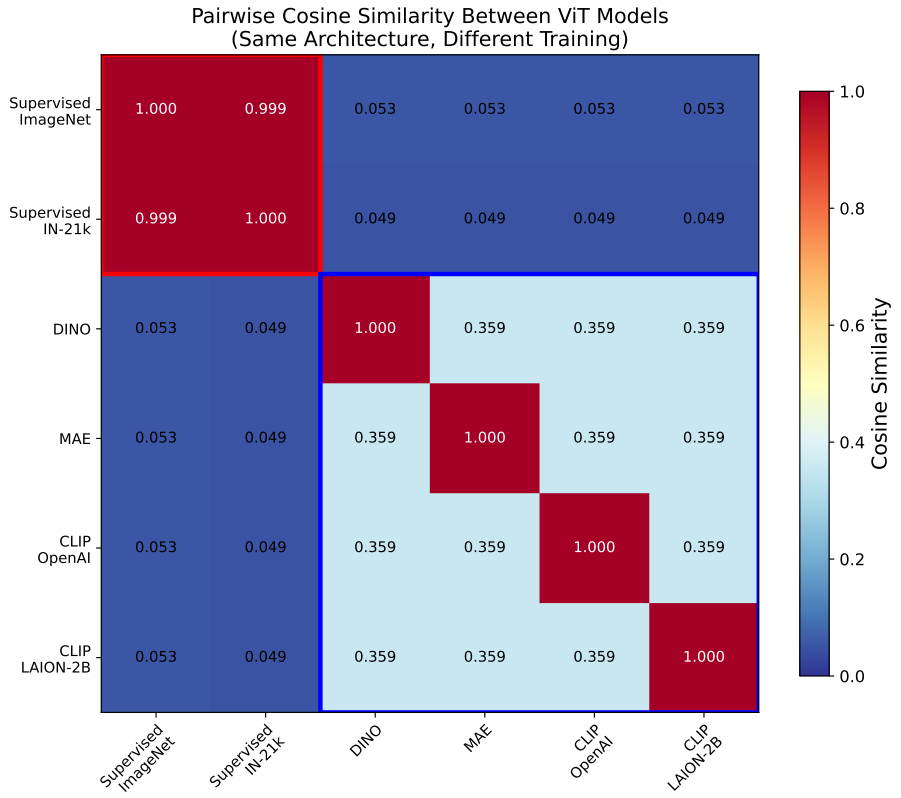

Figure 5: Pairwise cosine similarity between the 6 ViT models. Red box: supervised models are nearly identical ($\cos\theta \approx 0.999$). Blue box: self-supervised models are moderately similar to each other ($\cos\theta \approx 0.36$). Cross-method similarity is near zero ($\cos\theta \approx 0.05$).

Comparison	**Cosine Similarity**	**Interpretation**
Supervised vs. Supervised	0.999	Nearly identical
DINO vs. MAE vs. CLIP	$\sim$0.36	Moderately similar
Supervised vs. Self-supervised	$\sim$0.05	**Nearly orthogonal**

Table 3: Summary of cosine similarities between model groups.

Methodological Concern

Models with different training objectives are **nearly orthogonal** in weight space (cosine similarity ≈ 0.05), despite having identical architectures and learning from natural images. This directly contradicts the claim that models converge to a shared subspace "regardless of task."

4.5 Layer Contribution Analysis

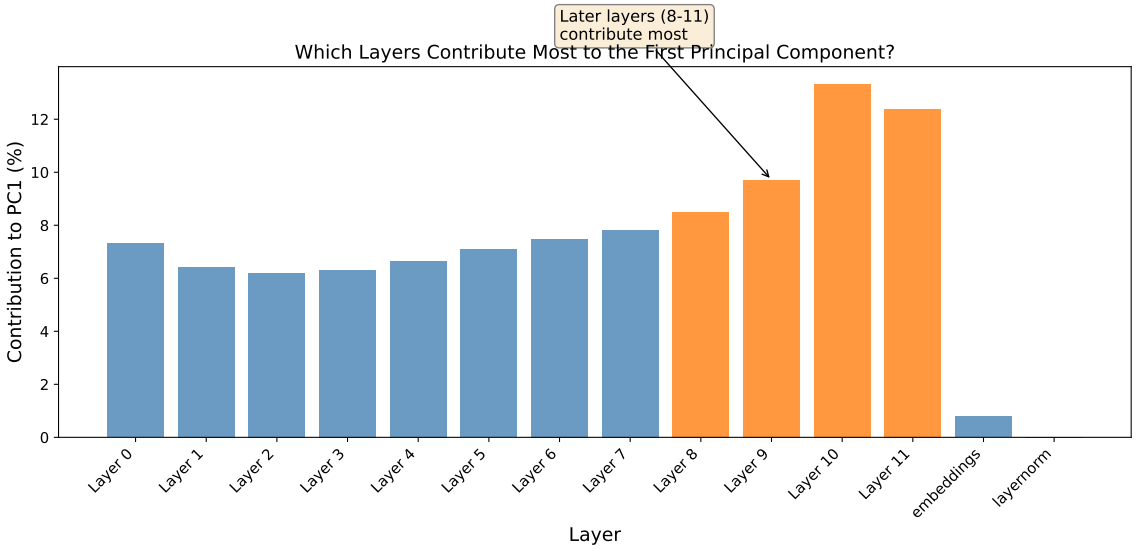

Figure 6: Contribution of each layer to PC1. Later layers (8–11) contribute most, suggesting PC1 captures task-specific representations rather than universal low-level features.

The fact that later layers contribute most to PC1 is consistent with our interpretation: PC1 represents the difference in high-level, task-specific representations between training objectives, not a universal feature of neural network learning.

5 Summary of Findings

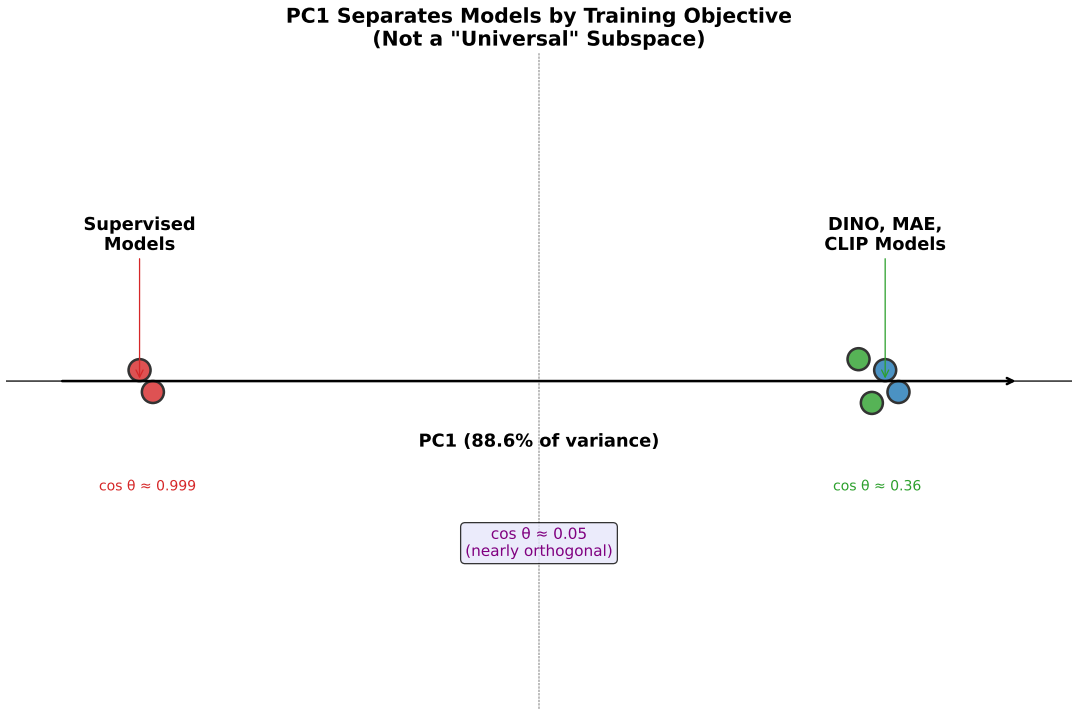

Figure 7: Schematic representation of our findings. PC1 separates models by training objective. Supervised and self-supervised models are nearly orthogonal despite identical architecture.

5.1 What We Found

1. **Apparent low-dimensional structure exists** when analyzing models from HuggingFace (PC1 explains 77–89% of variance).
2. **This structure reflects training methodology**, not universal convergence. Models cluster by training objective.
3. **Different training objectives produce nearly orthogonal solutions** (cosine similarity ≈ 0.05) despite:
   - Identical architecture (same weight space)
   - Same data domain (natural images)
   - Same optimization algorithm (gradient descent)
4. **The paper's ViT sample likely lacked diversity** in training objectives, leading to an apparent "universal" subspace that is actually the supervised ImageNet training subspace.

5.2 Reconciliation with the Original Paper

The paper's claim—"neural networks systematically converge to shared spectral subspaces regardless of initialization, task, or domain"—may be too strong. Our results suggest:

- **Within** a training methodology (e.g., all supervised ImageNet models), there may be a shared subspace.
- **Across** different training objectives, this universality breaks down.

The paper may have inadvertently sampled models that predominantly share the same training methodology, leading to a "universal" finding that doesn't generalize.

6 Conclusions

Conclusion

1. The "Universal Weight Subspace Hypothesis" is not supported when testing across genuinely diverse training objectives.
2. Models trained with supervised vs. self-supervised objectives occupy **nearly orthogonal** subspaces despite identical architectures.
3. The original paper's findings may reflect shared training ancestry (LoRA from same base model, ViTs fine-tuned from common checkpoints) rather than a fundamental property of neural network optimization.
4. Different loss functions create different loss landscapes, naturally leading to different solutions. This is expected behavior, not evidence against gradient-based learning.
5. Future work claiming "universal" properties of neural networks should ensure genuine diversity in:
   - Random initialization
   - Training objectives/loss functions
   - Training data
   - Model provenance (not fine-tuned from shared checkpoints)

7 Data and Reproducibility

7.1 Models Analyzed

All models were downloaded from HuggingFace using the `transformers` library:

```
from transformers import ViTForImageClassification
model = ViTForImageClassification.from_pretrained(model_name)
```

7.2 Key Numerical Results

20 HuggingFace ViTs:

- Variance in PC1: 77.0%
- k_{95}: 14/20 (70%)
- Effective dimension: 1.68
- Spectral ratio: 7.46

6 Models with Diverse Training:

- Variance in PC1: 88.6%
- Supervised vs. Supervised cosine: 0.999
- Self-supervised inter-method cosine: $\sim$0.36
- Supervised vs. Self-supervised cosine: $\sim$0.05

7.3 Code Availability

All analysis code is available in the accompanying repository:

- `replicate_vit_analysis_v4.py`: Main replication script
- `analyze_vit_pc1.py`: PC1 analysis with diverse training methods
- `check_vit_architectures.py`: Architecture verification
- `report/generate_figures.py`: Figure generation