

# Testing AI Vision's Understanding of High Fashion Nuances

*The first comprehensive evaluation of vision models for fashion intelligence*

## ABSTRACT

We present the first comprehensive benchmark evaluating vision models for fashion intelligence, testing CLIP, SigLIP, and DINOv2 on 12,147 Rick Owens runway images across 23 years of collections. Through 3.66 million image comparisons across three evaluation frameworks — impostor detection, collection cohesion, and exact matching — we demonstrate that SigLIP achieves superior fashion intelligence with +0.079 positive uncertainty detection gap and 63.5% collection purity, making it the only model suitable for production fashion AI systems despite 9.6x processing cost.

### Key Finding

**SigLIP** is the only model with proper uncertainty detection, achieving +0.079 positive noise gap while CLIP (−0.015) and DINOv2 (−0.070) show dangerous overconfidence.

## 1. METHODOLOGY

### 1.1 Model Selection

We evaluated three state-of-the-art vision models representing different architectural approaches to multi-modal understanding:

- **SigLIP** (google/siglip-so400m-patch14-384): Google's enhanced CLIP variant with sigmoid loss, 1152-dimensional embeddings, 400M parameters.
- **CLIP** (openai/clip-vit-base-patch32): OpenAI's original contrastive model, 512-dimensional embeddings.
- **DINOv2** (facebookresearch/dinov2\_vitb14): Meta's self-supervised vision model, 768-dimensional embeddings.

### 1.2 Dataset Construction

Our dataset comprises 12,147 Rick Owens runway images spanning 23 years (2002–2025) across multiple collections:

**Fall Collections:** 2002, 2003, 2004, 2008, 2013, 2015, 2016, 2021, 2024

**Spring Collections:** 2004, 2006, 2007, 2010, 2012, 2013, 2015, 2016, 2017, 2019, 2020, 2023, 2025

**Coverage:** Menswear, Ready-to-Wear, Beauty campaigns

### 1.3 Technical Infrastructure

Processing performed on M4 Mac with Apple Silicon GPU. Embeddings stored in Pinecone vector database for similarity search. All models implemented using PyTorch with transformers library.

## 2. THREE CORE INTELLIGENCE TESTS

| Test                      | Task                                                                                              | What good looks like                                                |
|---------------------------|---------------------------------------------------------------------------------------------------|---------------------------------------------------------------------|
| <b>Impostor Detection</b> | From pixels only, identify if candidates belong to the same designer and season/look as the query | Proper <i>uncertainty</i> on impostors — not overconfident matching |
| <b>Family Recognition</b> | Given one look from a season, find other looks from the same designer and season                  | High collection purity — low cross-season contamination             |
| <b>Needle Search</b>      | Find the exact same look in a database of 12,147 images                                           | Rank 1 match = exact designer, season, and look number              |

## 3. PERFORMANCE METRICS & TRADE-OFFS

### 3.1 Uncertainty Detection

Positive gap = model is properly less confident on impostors than true matches. Only SigLIP passes.

| Model         | Uncertainty Gap | Verdict | Interpretation                  |
|---------------|-----------------|---------|---------------------------------|
| <b>SigLIP</b> | <b>+0.079</b>   | PASS    | Proper uncertainty on impostors |
| <b>CLIP</b>   | −0.015          | FAIL    | Overconfident on noise          |
| <b>DINOv2</b> | −0.070          | FAIL    | Dangerously overconfident       |

### 3.2 Collection Purity

| Model         | Purity       | Note                            |
|---------------|--------------|---------------------------------|
| <b>SigLIP</b> | <b>63.5%</b> | Maintains season cohesion       |
| <b>CLIP</b>   | 48.8%        | Cross-season contamination      |
| <b>DINOv2</b> | 48.6%        | Near-random collection grouping |

### 3.3 Processing Speed

Designer Gauntlet Test (50 comparisons):

| Model         | Speed    | Trade-off                               |
|---------------|----------|-----------------------------------------|
| <b>CLIP</b>   | 1.9 min  | Fast but unreliable                     |
| <b>DINOv2</b> | 2.7 min  | Fast but unreliable                     |
| <b>SigLIP</b> | 18.7 min | 9.6x cost, only production-ready option |

## 4. TECHNICAL IMPLEMENTATION

### 4.1 SigLIP Embedding Generation

```
from transformers import AutoModel, AutoProcessor
import torch

# Load SigLIP model
siglip_model = AutoModel.from_pretrained(
    "google/siglip-so400m-patch14-384",
    torch_dtype=torch.float16
).to("mps") # Apple Silicon GPU

# Generate embedding
image = Image.open("rick-owens-fall-2011-menswear-39.jpg")
inputs = siglip_processor(images=image, return_tensors="pt")
with torch.no_grad():
    image_features = siglip_model.get_image_features(**inputs)
    embedding = image_features / image_features.norm(dim=-1, keepdim=True)
```

## 4.2 Vector Search Implementation

```
async def identify_image(self, image_embedding: List[float]):
    """Find most similar runway image using cosine similarity."""

    # Query Pinecone vector database
    query_result = self.img_index.query(
        vector=image_embedding,
        top_k=1,
        include_metadata=True,
        filter={"file_type": "look"} # Only runway looks
    )

    # Calculate confidence
    similarity_score = match.score
    if similarity_score >= 0.95:
        confidence = 99.0 # Near-perfect match
    elif similarity_score >= 0.90:
        confidence = 95.0 # Strong match
```

## 5. COMPLETE RESULTS SUMMARY

| Model         | Uncertainty Detection | Collection Purity | Needle Precision | Processing Speed    |
|---------------|-----------------------|-------------------|------------------|---------------------|
| <b>SigLIP</b> | +0.079 gap (PASS)     | 63.5%             | 100% accuracy    | 18.7 min (10x cost) |
| <b>CLIP</b>   | -0.015 gap (FAIL)     | 48.8%             | 90.0% accuracy   | 3.1 s               |
| <b>DINOv2</b> | -0.070 gap (FAIL)     | 48.6%             | 71.4% accuracy   | 4.2 s               |

## CONCLUSION

SigLIP is the only vision model suitable for production fashion AI. Its +0.079 uncertainty gap and 63.5% collection purity demonstrate genuine semantic understanding of designer aesthetic, not just visual similarity matching. CLIP and DINOv2's negative uncertainty gaps — meaning they are *more* confident on impostors than true matches — make them categorically unsuitable for high-precision fashion retrieval.

The 9.6x processing cost of SigLIP is a real constraint but acceptable at production scale for high-value fashion intelligence applications. The benchmark establishes SigLIP at [google/siglip-so400m-patch14-384](https://github.com/google/siglip-so400m-patch14-384) as the foundational model for any serious fashion AI system requiring reliable similarity search across large runway archives.