VLM DNA: TRACING VISION-LANGUAGE MODELS VIA DISTANCE-PRESERVING REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying vision-language models (VLMs) and tracing their relationships across fine-tuning trajectories requires compact model representations that can be extracted without access to model internals. We study training-free VLM identity representation extraction using only image-question queries and generated responses, without modifying model parameters or requiring shared architectures. Existing text-only response representations do not explicitly account for the relationship between generated answers and their visual inputs, potentially overlooking useful signals for distinguishing VLMs. We introduce VLM DNA, a compact representation of model behavior, and DNAExtract, a training-free framework for extracting it. DNAExtract probes each model with a shared set of image-question pairs and encodes the images and generated responses using frozen CLIP image and text encoders. It then constructs outer-product interaction matrices to represent image-response associations and applies sparse random projection to obtain compact DNA vectors designed to approximately preserve pairwise distances between the original cross-modal representations. Experiments on 39 open-source VLMs from seven families demonstrate that VLM DNA achieves an AUC of 0.929 for relation detection, an AUC of 0.923 for one-to-one verification, and a rank-1 accuracy of 89.38% with a mean average precision of 0.936 for one-to-many identification. These results outperform text-only response embedding baselines and support the utility of cross-modal representations for VLM identification and tracing. Code will be released upon acceptance.

1 INTRODUCTION

Vision-language models (VLMs) are commonly developed by adapting existing checkpoints through transfer learning, knowledge distillation, and visual instruction tuning (Zhang et al., 2024; Liu et al., 2023). These practices produce families of related models whose provenance can be difficult to establish. An analysis of recorded lineage fields on Hugging Face reports that approximately 60% of VLM releases in the analyzed dataset have no recorded parent (Cebrian, 2026). Incomplete provenance complicates both security auditing and attribution: prior work shows that backdoors can transfer from teacher to student models through distillation (Cheng et al., 2024), while undocumented derivation creates challenges for assessing license compliance (Duan et al., 2025). Metadata provides useful evidence when available, but may be incomplete or difficult to verify, particularly for models accessed only through APIs. These limitations motivate behavioral methods for investigating model relationships using query access alone.

Several lines of research provide relevant tools, but target different objectives. VLM routing and model selection focus on choosing a suitable model for a given query (Huang et al., 2026; Cheng & Dong, 2026). Ownership verification typically assesses whether a suspect model is derived from a particular reference model (Wang et al., 2025; Zhao et al., 2026; Xia et al., 2026; Zheng et al., 2026b;a). These objectives do not by themselves establish a common representation for comparing heterogeneous VLMs across multiple tracing tasks. Lineage analysis more directly studies model relationships, including through text-only behavior (Yax et al., 2025; Wu et al., 2026). Our goal is to construct an identity representation for VLMs without accessing their weights, modifying their parameters, or requiring shared architectures.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Extending behavioral model representations to VLMs introduces a central design question: how should visual inputs and generated responses jointly contribute to model comparison? Text-only response embeddings summarize answer semantics but do not explicitly represent the association between an answer and its corresponding image. We hypothesize that responses to a shared set of visual queries contain behavioral signatures that are informative about model identity and relatedness. This leads to our research question: *Can we extract compact representations from VLM query responses that support relation detection, verification, and base-model identification within a common space?*

We introduce VLM DNA, a compact VLM behavior representation, and DNAExtract, a training-free extraction framework. Building on the functional view of LLM DNA (Wu et al., 2026), we represent behavior through responses to paired image–question inputs. DNAExtract queries each model with the same probe set as a common basis for comparison, encodes each image and generated response with frozen CLIP image and text encoders (Radford et al., 2021), and forms their outer product as a probe-level interaction representation. It concatenates these representations and applies a shared sparse random projection (Achlioptas, 2003; Li et al., 2006) to obtain a compact DNA vector that approximately preserves pairwise distances while reducing storage and comparison costs. With probes, encoders, and projection fixed, additional models can be represented without retraining.

We evaluate VLM DNA on 39 open-weight VLMs from seven families using 600 shared probes from six multimodal benchmarks. The representations support relation detection (AUC 0.929), one-to-one verification (AUC 0.923), and one-to-many base-model identification (Rank-1 89.38%; mean average precision (mAP) 0.936), outperforming text-only response embedding baselines. We further examine distance preservation, inter-model distance stability across disjoint probe sets, and the effects of individual representation components. Section 5 provides detailed results.

Contributions. (1) We formulate black-box VLM tracing as the extraction of a common behavioral representation that supports relation detection, verification, and base-model identification within a unified identity space. (2) We introduce DNAExtract, a training-free framework that constructs image–response interaction representations and compresses them into compact VLM DNA vectors through a shared sparse random projection. (3) We evaluate the framework on 39 VLMs from seven families, examining its tracing performance, distance preservation, and sensitivity to probe selection and representation design.

2 Related Work

2.1 Model Identification, Ownership Verification, and Lineage

Model identification and ownership verification. Model identification seeks to determine which model produces observed outputs, whereas ownership verification assesses whether a suspect model is related to a particular protected model. For text-only LLMs, existing approaches use strategic queries, learned text fingerprint detectors, and representation-based fingerprints (Pasquini et al., 2025; Fu et al., 2025; Zhang et al., 2025). Other methods examine model internals to investigate provenance or independence (Nikolic et al., 2025; Zhu et al., 2025). For VLMs, PLA constructs adversarial images for copyright tracking without modifying the released model (Wang et al., 2025). SIF optimizes semantically coherent fingerprint queries (Zhao et al., 2026), while AGDI combines response-level and CLIP-based semantic objectives with auxiliary adversarial training (Xia et al., 2026). UMMF constructs fingerprints through unlearning-induced differences in memorization (Zheng et al., 2026b), and SALSF embeds fingerprint signals through synonym-aware logit shaping (Zheng et al., 2026a). These approaches differ in their construction procedures and access requirements, even when verification is performed against a black-box suspect model. Our focus is on extracting model representations with shared probes and a fixed mapping for comparison across multiple tracing tasks.

Model provenance and phylogeny. Lineage analysis investigates relationships among models beyond individual ownership claims. Neural Lineage studies parent-model identification and proposes both learning-free and learning-based approaches (Yu & Wang, 2024). Model DNA learns compact representations from training data and input–output information for provenance identification across vision and language tasks (Mu et al., 2024). For text-only LLMs, PhyloLM estimates behavioral similarity from sampled generations, using character-based comparisons to accommodate

different tokenizers (Yax et al., 2025). LLM DNA develops a training-free functional representation framework with a bi-Lipschitz formulation and an existence result for finite model collections (Wu et al., 2026). RAFP constructs rare-region prompt fingerprints through gradient-based optimization and evaluates their persistence under downstream adaptations (Tsai et al., 2025). Complementary white-box approaches analyze centered residual signatures and spectral structure in model weights (Thakur & Khoury, 2026; Chen et al., 2026). Model Provenance Set (MPS) applies sequential statistical testing to construct candidate provenance sets with asymptotic coverage guarantees under its assumptions (Qiu et al., 2026). These approaches provide different forms of behavioral, structural, and statistical evidence. We study compact representations of image-conditioned VLM responses for relation detection, verification, and base-model identification.

2.2 Cross-Modal Representations and Distance Preservation

Representing image-response associations. Vision–language pretraining provides shared feature spaces for comparing images and text. CLIP learns image–text alignment through a contrastive objective (Radford et al., 2021), while SigLIP uses a sigmoid loss (Zhai et al., 2023). BLIP and LLaVA further develop multimodal understanding and generation through bootstrapped pretraining and visual instruction tuning, respectively (Li et al., 2022; Liu et al., 2023). These methods provide representations and capabilities for vision–language tasks. In our DNAExtract , frozen CLIP encoders serve as a common external feature extractor for images and responses produced by heterogeneous VLMs. The resulting features characterize observable image–response associations; their usefulness for model tracing is evaluated empirically.

Distance-preserving compression. The Johnson–Lindenstrauss lemma establishes that a finite set of Euclidean vectors admits a low-dimensional embedding with bounded pairwise distance distortion (Johnson & Lindenstrauss, 1984). Sparse random projections offer computationally efficient constructions, with accuracy depending on the projection distribution, sparsity, and target dimension (Achlioptas, 2003; Li et al., 2006). These results concern distances between the vectors being projected. They do not establish that a chosen representation separates model lineages or uniquely identifies model ancestry. Building on the functional representation perspective of LLM DNA (Wu et al., 2026), DNAExtract applies sparse random projection to representations constructed from shared image–question probes. We evaluate distance preservation and tracing performance separately. Table 1 summarizes the construction procedures and evidence used by representative methods.

Table 1: Comparison of representative model evolution and tracing methods from the model provenance and phylogeny literature. (✓: supported; ✗: not supported in the reported setting; –: not established in this comparison.)

Method	Objective	Target models	Black-box access	Training-free	No model modification	Cross-modal behavior
Neural Lineage (Yu & Wang, 2024)	Model lineage	General NN	✗	–	✓	✗
Model DNA (Mu et al., 2024)	Model provenance	Vision/language	–	✗	–	–
PhyloLM (Yax et al., 2025)	Phylogeny	LLMs	✓	–	✓	✗
LLM DNA (Wu et al., 2026)	Model lineage	LLMs	✓	✓	✓	✗
RAFP (Tsai et al., 2025)	Model lineage	LLMs	✓	–	✓	✗
Training Leaves Traces (Thakur & Khoury, 2026)	Model lineage	LLMs	✗	✓	✓	✗
Who Built This Model? (Chen et al., 2026)	Model lineage	LLMs	✗	✓	✓	✗
Provable MPS (Qiu et al., 2026)	Model provenance	LLMs	✓	✓	✓	✗
VLM DNA (Ours)	**Model lineage**	**VLMs**	✓	✓	✓	✓

3 Preliminaries and Motivation

3.1 Preliminaries

LLM DNA (Wu et al., 2026) represents the functional behavior of a large language model (LLM) as a compact, low-dimensional vector. Given two LLMs f_1 and f_2, their DNA representations $\boldsymbol{\tau}_{f_1}, \boldsymbol{\tau}_{f_2} \in \mathbb{R}^L$ satisfy the bi-Lipschitz condition:

$$c_1 d_{\mathcal{H}}(f_1, f_2) \leq \|\boldsymbol{\tau}_{f_1} - \boldsymbol{\tau}_{f_2}\|_2 \leq c_2 d_{\mathcal{H}}(f_1, f_2), \tag{1}$$

where $d_{\mathcal{H}}$ denotes the functional distance between models and $0 < c_1 < c_2$ are constants. This condition establishes two fundamental properties: inheritance and genetic determinism. In practice,

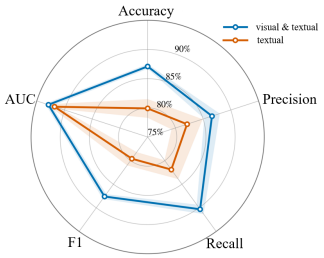

Figure 1: VLM relation detection using textual-only MPNet representations and joint visual–textual representations from VLM DNA.

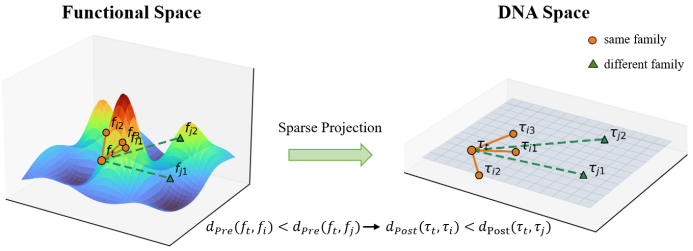

Figure 2: Conceptual illustration of the distance-preservation objective. A shared sparse projection compresses probe-based behavioral representations E_f into DNA vectors $\tau_f = RE_f$, aiming to approximately preserve their pairwise Euclidean distances.

LLM DNA is extracted by encoding model responses to a shared set of text prompts into semantic embeddings, concatenating these embeddings, and applying Gaussian random projection to obtain a low-dimensional DNA vector.

3.2 Motivation

We first examine whether textual-only representations are sufficient for detecting relationships between VLMs. As shown in Figure 1, VLM DNA outperforms the textual-only MPNet baseline in the considered relation detection setting. Although textual responses reflect the observable behavior of VLMs, they are generated in response to specific visual inputs and textual instructions. Encoding textual responses alone does not explicitly retain their association with the corresponding visual context. The observed performance gap therefore motivates us to investigate a multimodal behavioral representation that jointly encodes images and their corresponding responses, aiming to characterize VLM functional behavior under shared visual and textual inputs.

Beyond capturing multimodal behavior, extending DNA to VLMs requires retaining the fundamental properties of inheritance and genetic determinism (Wu et al., 2026). These properties call for a compact representation that preserves behavioral relationships between models: VLMs with similar behavior should have similar DNA representations, while nearby DNA vectors should indicate similar represented behavior. However, compressing high-dimensional multimodal representations into compact vectors may distort inter-model distances and undermine these properties. As illustrated in Figure 2, we therefore aim to construct a compact VLM DNA representation that characterizes image-conditioned functional behavior while approximately preserving pairwise behavioral distances between models.

4 Methodology

4.1 VLM DNA

We study VLM lineage tracing through functional behavior under black-box access. Following LLM DNA (Wu et al., 2026), we define model identity through functional similarity rather than parameters, tokenizers, or architectures. We extend this perspective to joint image–text inputs while retaining the inheritance and genetic-determinism properties of the DNA formulation.

Definition 1 (Vision–Language Model)**.** *Let $\mathcal{X}$ be a finite set of digital images with bounded dimensions and finite pixel precision. Let $\mathcal{T}$ and $\mathcal{Y}$ be finite sets of input and output strings over a finite alphabet with bounded lengths. Under a specified decoding and truncation protocol, a text-generative VLM induces a function $f : \mathcal{X} \times \mathcal{T} \to \Delta^{|\mathcal{Y}|-1}$, where $f(x,t)_y$ is the probability of generating response $y \in \mathcal{Y}$. The collection of these functions is denoted by $\mathcal{F}$.*

Let μ be a probability distribution with full support on $\mathcal{X} \times \mathcal{T}$. We define the functional distance between two VLMs as

$$d_H(f_a, f_b) = \left(\sum_{(x,t)\in\mathcal{X}\times\mathcal{T}} \mu(x,t) \, \|f_a(x,t) - f_b(x,t)\|_2^2 \right)^{1/2}. \tag{2}$$

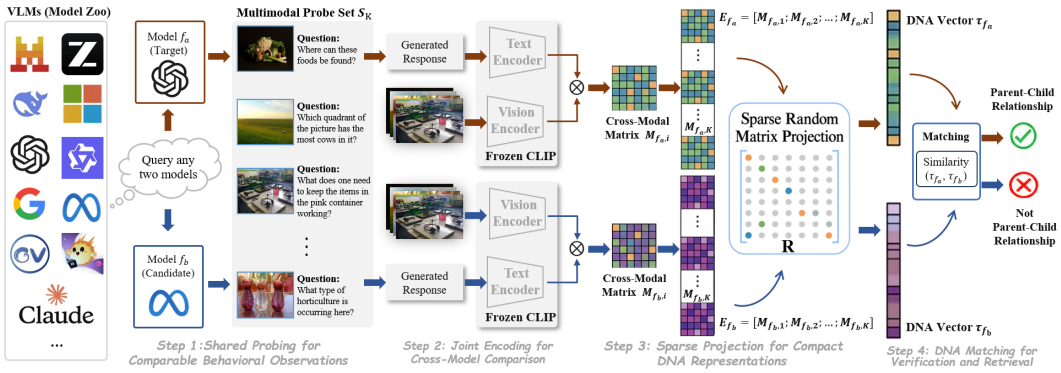

Figure 3: Overview of DNAExtract. Step 1 queries each VLM f with a shared set $\mathcal{S}_K$. Step 2 encodes each image and response using frozen CLIP encoders ϕ_V and ϕ_T, forms cross-model matrices $\mathbf{M}_{f,i}$, and concatenates them into $\mathbf{E}_f$. Step 3 applies a shared sparse projection $\mathbf{R}$ to obtain the DNA vector τ_f. Step 4 uses DNA distances for 1:1 verification and closed-set 1:N base-modal identification.

This distance compares response distributions over common strings, independently of model-specific tokenization. Finiteness is a theoretical device; the extraction pipeline neither enumerates these spaces nor accesses complete response distributions.

Definition 2 (VLM DNA Representation)**.** *For a finite model collection $\mathcal{F}_M = \{f_1, \ldots, f_M\} \subset \mathcal{F}$, a VLM DNA representation assigns each model a vector $\tau_f = g(f) \in \mathbb{R}^k$. The mapping satisfies a bi-Lipschitz condition on this collection:*

$$c_1 d_H(f_a, f_b) \leq \|\tau_{f_a} - \tau_{f_b}\|_2 \leq c_2 d_H(f_a, f_b), \tag{3}$$

where $0 < c_1 \leq c_2$ are shared across all model pairs in $\mathcal{F}_M$.

This definition retains the two DNA properties introduced by Wu et al. (2026).

Proposition 1 (Properties of VLM DNA)**.** *The VLM DNA representation defined above satisfies:* (i) Inheritance. *If an adaptation produces f' from f with $d_H(f, f') \leq \eta$, then $\|\tau_f - \tau_{f'}\|_2 \leq c_2\eta$. Thus, adaptations that induce small functional changes produce nearby DNA representations;* (ii) Genetic determinism. *If two models satisfy $\|\tau_{f_a} - \tau_{f_b}\|_2 \leq \delta$, then $d_H(f_a, f_b) \leq \delta/c_1$. Nearby DNA representations therefore imply similar functional behavior under μ. This implication concerns behavioral similarity rather than proof of shared ancestry.*

These properties follow from the upper and lower bounds in Equation 3. Because the input and output spaces are finite, each model admits a weighted representation in a finite-dimensional Euclidean space. The finite-set existence argument of LLM DNA therefore extends to the multimodal setting through the Johnson–Lindenstrauss lemma (Johnson & Lindenstrauss, 1984; Wu et al., 2026). This theoretical definition specifies the representation objective. Our practical pipeline uses finite probes and frozen encoders to construct a computable behavioral representation. We distinguish distances in this representation from d_H and assess compression fidelity and lineage utility separately.

4.2 DNAExtract: VLM DNA Extraction

We introduce **DNAExtract**, a training-free framework for extracting VLM DNA from black-box multimodal behavior. To operationalize the preceding representation objective, we (i) characterize each VLM over a tractable set of shared multimodal probes and (ii) compress the resulting high-dimensional representation into a compact DNA vector. Figure 3 summarizes extraction and matching in four stages: shared probing, joint image–response encoding, sparse projection, and DNA matching. Steps 1–3 constitute DNAExtract and are summarized in Algorithm 1 (Appendix A.2). Extraction requires only query access to the target VLM and frozen CLIP encoders, without additional training or model modification.

Step 1: Shared Probing for Comparable Behavioral Observations. Because evaluating d_H over the full input space is computationally prohibitive, we follow the probing strategy of Wu et al. (2026) and construct a computable surrogate from a finite set of multimodal inputs.

Definition 3 (Multimodal Stochastic Functional Distance)**.** *Let μ be a probability distribution over $\mathcal{X} \times \mathcal{T}$, and let $\mathcal{S}_K = \{(x_i, t_i)\}_{i=1}^K$ be K i.i.d. probes drawn from μ. Let $\hat{y}(f, x, t; \xi)$ denote the response generated by VLM f under a fixed decoding configuration, where ξ represents generation randomness. The joint image–response representation is*

$$\rho(f, x, t; \xi) = \mathrm{vec}\Big(\phi_V(x)\, \phi_T(\hat{y}(f, x, t; \xi))^\top\Big) \in \mathbb{R}^{D^2}, \tag{4}$$

where ϕ_V and ϕ_T are frozen CLIP visual and text encoders with output dimension D. We define the stochastic representation distance as

$$d_f(f_1, f_2) = \mathbb{E}_{\mathcal{S}_K, \boldsymbol{\xi}} \left[\left(\sum_{i=1}^K \|\rho(f_1, x_i, t_i; \xi_{1,i}) - \rho(f_2, x_i, t_i; \xi_{2,i})\|_2^2 \right)^{1/2} \right], \tag{5}$$

where the expectation is over probe sampling and response-generation randomness under the specified sampling protocol.

In practice, we use a fixed shared probe set with equal contributions from six benchmarks and generate a response for each model and probe. For the realized responses, the empirical distance is

$$d_{\mathrm{pre}}(f_1, f_2; \mathcal{S}_K) = \left(\sum_{i=1}^K \|\rho(f_1, x_i, t_i) - \rho(f_2, x_i, t_i)\|_2^2 \right)^{1/2}, \tag{6}$$

where generation randomness is omitted from the notation. Shared inputs and ordering provide matched observation conditions. The visual feature alone is identical across models for a given probe; model-dependent variation enters through the generated response. This empirical distance concerns the encoded observations and is not assumed to be an isometric estimate of d_H.

Step 2: Joint Encoding for Cross-Model Comparison. We use frozen CLIP encoders ϕ_V and ϕ_T to embed each image–response pair: $\mathbf{v}_i = \phi_V(x_i) \in \mathbb{R}^D$ and $\mathbf{s}_{f,i} = \phi_T(\hat{y}_{f,i}) \in \mathbb{R}^D$. The same encoders and feature processing establish common coordinates across heterogeneous VLMs. We form the outer product $\mathbf{M}_{f,i} = \mathbf{v}_i \mathbf{s}_{f,i}^\top \in \mathbb{R}^{D \times D}$, whose (m, n)-th entry is the product of the m-th visual coordinate and the n-th textual coordinate.

Flattening and concatenating all K matrices gives

$$\mathbf{E}_f = [\mathrm{vec}(\mathbf{M}_{f,1}); \ldots; \mathrm{vec}(\mathbf{M}_{f,K})] \in \mathbb{R}^{KD^2}. \tag{7}$$

This construction retains the shared probe ordering before projection. Each $\mathrm{vec}(\mathbf{M}_{f,i})$ is the per-probe representation in Definition 3, and $d_{\mathrm{pre}}(f_a, f_b) = \|\mathbf{E}_{f_a} - \mathbf{E}_{f_b}\|_2$.

Step 3: Sparse Projection for Compact DNA Representations. The joint representation has dimension $KD^2 \approx 3.5 \times 10^8$ under our experimental configuration. Compared with concatenated D-dimensional response embeddings, the outer product increases this dimension by a factor of D. With $k = 512$, a dense projection matrix would contain approximately 1.81×10^{11} entries, requiring roughly 724 GB in float32. We therefore adopt a sparse random projection (Achlioptas, 2003; Li et al., 2006), detailed in Appendix A.1.

The matrix $\mathbf{R} \in \mathbb{R}^{k \times KD^2}$ is sampled once per extraction configuration and shared across all models. The extracted DNA vector is $\tau_f = \mathbf{R}\mathbf{E}_f \in \mathbb{R}^k$. With the normalization specified in Appendix A.1, the construction preserves squared Euclidean distances in expectation:

$$\mathbb{E}_{\mathbf{R}} \left[\|\tau_{f_a} - \tau_{f_b}\|_2^2 \right] = \|\mathbf{E}_{f_a} - \mathbf{E}_{f_b}\|_2^2. \tag{8}$$

Step 4: DNA Matching for Verification and Retrieval. We use the extracted DNA vectors for two matching tasks. For 1:1 verification, we predict that two VLMs f_a and f_b share the same base-model identity when $\|\tau_{f_a} - \tau_{f_b}\|_2 \leq \theta$. For 1:$N$ identification, let $\mathcal{G} = \{g_1, \ldots, g_M\}$ be a gallery of known base models. Given a query model f_q, we retrieve $\hat{g} = \arg\min_{g \in \mathcal{G}} \|\tau_{f_q} - \tau_g\|_2$. We report verification and retrieval metrics in Section 5.

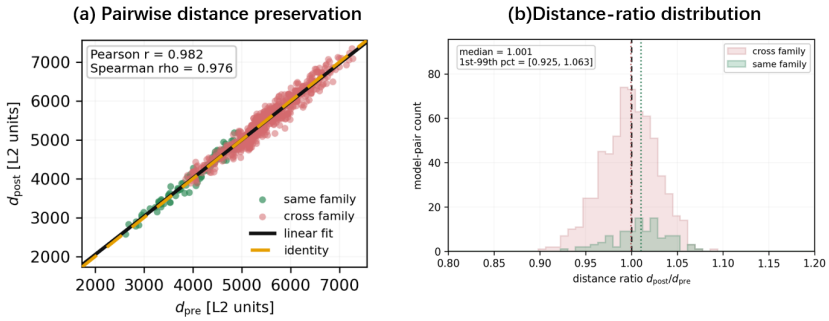

Figure 4: Distance preservation under sparse projection. (a) DNA distances versus pre-projection distances for model pairs, with the identity line indicating exact preservation. (b) distance-ratio distributions for same-family and cross-family pairs; a ratio of one indicates unchanged distance.

5 Experiments

We evaluate the VLM DNA representations extracted by DNAExtract along two dimensions: distance preservation under sparse projection and three identity tasks: 1:1 verification, 1:N identification, and relation detection. We additionally examine consistency across disjoint probe sets and representation drift under controlled fine-tuning.

5.1 Experimental Setup

Model pool. We collect 39 open-weight VLMs from Hugging Face, grouped into seven families. Each family contains one designated base checkpoint and fine-tuned variants spanning general instruction tuning, safety alignment, document understanding and OCR, chart and table understanding, medical and domain specialization, preference optimization, and style transfer. Appendix B.1 lists model identifiers, roles, and parameter counts. We assign same-family and cross-family labels using the Hugging Face model tree; cross-family does not imply independent development or the absence of shared components.

Probe set. We construct a shared set $\mathcal{S}_K$ of $K = 600$ image–question pairs, with 100 randomly selected examples from each of six benchmarks: ScienceQA for science reasoning with diagrams (Lu et al., 2022), A-OKVQA for external-knowledge VQA (Schwenk et al., 2022), ChartQA for chart interpretation (Masry et al., 2022), MMMU for college-level multimodal reasoning (Yue et al., 2024), MathVista for mathematical reasoning over figures (Lu et al., 2024), and DocVQA for document understanding (Mathew et al., 2021). Multiple-choice options are appended to questions in a uniform format. All models receive the same probe content.

Implementation details. DNAExtract uses shared probes, frozen encoders, and a shared projection matrix to produce 512-dimensional representations without training on model identities or family labels. Verification and retrieval use Euclidean distances directly, whereas relation detection trains an SVM on the extracted representations. The repeated downstream evaluations use five projection seeds. We set $k = 512$ and $D = 768$. Generation uses each model's native chat template, a maximum generation length of 1024 tokens, temperature 0.7, and top-p 0.9. The supplementary dimension ablation shows that downstream performance approaches a plateau at moderate dimensions, motivating the default k.

5.2 Distance Preservation of Sparse Projection

We first directly test whether the sparse projection preserves model-pair geometry. For each of the unordered pairs, we compare its distance in the full functional representation, $d_{\text{pre}}(a, b) = \|\mathbf{E}_{f_a} - \mathbf{E}_{f_b}\|_2$, with its distance in the DNA space, $d_{\text{post}}(a, b) = \|\tau_{f_a} - \tau_{f_b}\|_2$. As shown in Figure 4, the pairwise distances follow the identity line, with Pearson $r = 0.982$ and Spearman $\rho = 0.976$. The distance ratio $d_{\text{post}}/d_{\text{pre}}$ has median 1.001 and 1st–99th percentile range $[0.925, 1.063]$, while same-family and cross-family pairs have similar medians (1.010 and 0.999). These results demonstrate that sparse projection approximately preserves the pairwise distances that define their functional-space geometry.

Table 2: Family-level LOFO relation detection. Values are mean ± standard deviation over five seeds after averaging across held-out families. Higher values are better. **Bold** and underlining indicate the highest and second-highest mean, respectively.

Method	Acc.	Prec.	Rec.	F1	AUC
Random	0.475 ± 0.028	0.473 ± 0.024	0.479 ± 0.026	0.471 ± 0.017	0.481 ± 0.065
MPNet + SVM	0.799 ± 0.016	0.821 ± 0.028	0.819 ± 0.020	0.796 ± 0.014	0.918 ± 0.014
VLM DNA + SVM	$\mathbf{0.871 \pm 0.004}$	$\mathbf{0.866 \pm 0.014}$	$\mathbf{0.903 \pm 0.010}$	$\mathbf{0.876 \pm 0.005}$	$\mathbf{0.929 \pm 0.006}$

Table 3: Fusion ablation under family-level LOFO. Values are mean ± standard deviation over five projection seeds. The CLIP encoders and evaluation protocol are shared across variants. Higher values are better; **bold** and underlining mark the highest and second-highest mean, respectively.

Representation	Acc.	Prec.	Rec.	F1	AUC
Concat	0.859 ± 0.004	0.863 ± 0.011	0.878 ± 0.013	0.860 ± 0.006	0.920 ± 0.013
Hadamard	0.818 ± 0.015	0.831 ± 0.027	0.807 ± 0.006	0.805 ± 0.010	0.866 ± 0.011
Outer product (VLM DNA)	$\mathbf{0.871 \pm 0.004}$	$\mathbf{0.866 \pm 0.014}$	$\mathbf{0.903 \pm 0.010}$	$\mathbf{0.876 \pm 0.005}$	$\mathbf{0.929 \pm 0.006}$

5.3 Detecting VLM Relationships

Test whether VLM DNA separates same-family from cross-family pairs. On 741 unordered pairs (102 same-family positives, 639 cross-family negatives), an RBF-kernel SVM classifies each pair's absolute DNA difference $x_{ab} = |\tau_{f_a} - \tau_{f_b}| \in \mathbb{R}^{512}$. Use seven-fold LOFO with held-out families excluded from training, their pairs as the candidate test set, and train/test balanced at $1:1$; average metrics across held-out families. Compare with a response-only MPNet baseline (`sentence-transformers/all-mpnet-base-v2`) and uniform random guessing under matched folds, seeds, and SVM settings. Table 2 shows VLM DNA leading all metrics (AUC 0.929 ± 0.006), improving MPNet accuracy from 0.799 to 0.871 and F1 from 0.796 to 0.876, The result indicate that cross-modal behavior captures lineage information beyond response text.

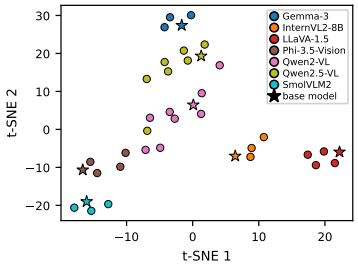

Figure 5: t-SNE visualization of VLM DNA representations.

Fusion ablation. Using the same CLIP encoders, output dimension, projection scheme, classifier settings, and LOFO protocol, we compare the outer product, concatenation, and the Hadamard product. The outer product leads every metric (Table 3; AUC 0.929 versus 0.920 and 0.866), supporting its use for relation detection in DNAExtract. Figure 5 provides a complementary visualization of family-associated structure in the DNA representations.

5.4 1:1 Identity Verification

We next evaluate whether Euclidean distance between DNA vectors supports threshold-based verification. Using 102 same-family genuine pairs and 639 cross-family impostors, sweep an L2 threshold and report EER, ROC AUC, and d'; chance yields AUC 0.5 and EER 50%. Across five projection seeds, VLM DNA obtains AUC 0.923 ± 0.007, EER $13.94\% \pm 0.67\%$, and $d' = 1.98 \pm 0.05$, above random. Figure 6 shows that the left panel pools L2 distances over these seeds, density-normalized for unequal counts; genuine distances cluster lower and impostors higher, the dashed line marks the mean EER threshold, and overlap contains boundary pairs no threshold separates due to fine-tuning drift or cross-family similarity. The right panel shows the five-seed ROC curve. Thus, VLM DNA encodes family-level identity in its distance geometry and supports threshold-based verification without training a classifier.

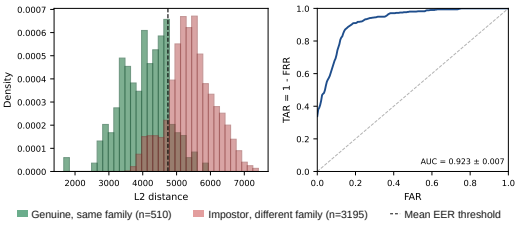

Figure 6: 1:1 identity verification over the five sparse-projection seeds. Left: pooled L2-distance distributions of genuine and impostor model pairs. Right: ROC curve averaged over the five seeds.

5.5 1:N Identity Identification

In the 1:N task, the gallery contains seven official base models, and 32 fine-tuned variants act as unknown probes. Assign each probe the nearest gallery model by L2 distance, and report Rank-1 accuracy, mean average precision (mAP), and cumulative match characteristic (CMC); chance Rank-1 is $1/7 \approx 14.3\%$. VLM DNA achieves Rank-1 $89.38\% \pm 2.80\%$ and mAP 0.936 ± 0.015 over five projection seeds. Figure 7 shows the five-seed-averaged CMC curve, which exceeds chance at every rank; Appendix C.1 reports per-family results. Thus, VLM DNA retrieves the correct base model for most fine-tuned variants, indicating that family-level identity remains recoverable after fine-tuning.

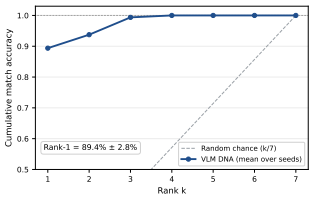

Figure 7: Cumulative matching curve of 1:N identification ($k/7$).

5.6 Cross-Probe Consistency

We examine how the choice of the sampled probe distribution $\mathcal{S}_K$ affects the quality of the extracted VLM DNAs. We use a Mantel test between DNAs from two disjoint sets: 100 probes from ScienceQA plus 100 from A-OKVQA, and 100 from ChartQA plus 100 from DocVQA. We compute DNAs for the same 39 models, calculate distances for all 741 model pairs, visualize both distance structures in Figure 8, and compute the Pearson and Spearman correlations with permutation-based p-values. We observe a strong correlation (Pearson $r = 0.852$; Spearman $r = 0.808$; $p < 0.001$). These results show that VLM DNA structure does not depend on the sampled probe distribution, supporting the stability of our extraction approach.

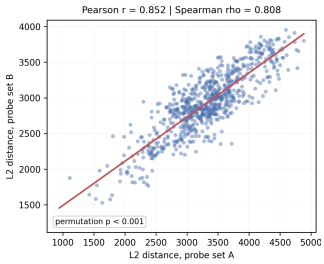

Figure 8: Mantel test across two disjoint 200-probe sets; the red line shows the linear least-squares fit.

5.7 Controlled Fine-Tuning Drift

For controlled adaptation, we fully fine-tune SmolVLM2-2.2B-Instruct on nested subsets of the 113K PRISM demonstrations (10/100/1,000/10,000 samples; (Wang et al., 2026)), which postdate its public training sources. We fix the full-parameter protocol and vary only sample count; DNA uses the main experiments' 600 probes and five projection seeds. Mean L2 drift across 10/100/1,000/10,000 is 2761.96/2716.90/2931.51/3292.54, with normalized drift from 0.3023 to 0.3603. The first two values remain within seed variation, whereas the last two increase consistently. Figure 9 shows five-seed-averaged drift across 128 blocks of 512 dimensions; most retain their relative intensity with only local changes, indicating that fine-tuning alters local traits but preserves inherited behavior. At 10k samples, L2 drift remains below the 1:1 EER threshold, keeping the model on the same side of the same-family boundary.

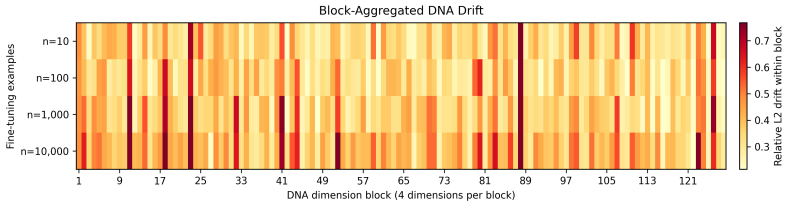

Figure 9: Shift of DNA values after full-parameter fine-tuning SmolVLM2-2.2B-Instruct with increasing amounts of external data.

6 Conclusion

We introduced VLM DNA, a compact behavioral identity representation, and DNAExtract, its training-free cross-modal extractor, which uses shared image–question probes, frozen CLIP encoders, and sparse random projection. These distance-preserving vectors support 1:1 verification, 1:N identification, and relation detection; on 39 VLMs from seven families, they preserve pairwise distances, remain stable across disjoint probes, and keep fine-tuned variants near their base models.

7 AI USE STATEMENT

We used generative AI tools to assist with language polishing, literature retrieval and organization, research ideation, and the drafting of selected portions of the manuscript. We did not use these tools to generate experimental data, perform mathematical proofs, or replace the authors' responsibility for the scientific analysis and conclusions. All AI-assisted content was carefully reviewed and verified by the authors. We independently checked the technical descriptions, experimental settings, implementation details, and reported results. We take full responsibility for the final content of this work, including all claims, analyses, and artifacts.

REFERENCES

Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003.

Manuel Cebrian. Founder effects shape the evolutionary dynamics of multimodality in open LLM families. *arXiv preprint arXiv:2603.22287*, 2026.

Yiwei Chen, Bingqi Shang, and Sijia Liu. Who built this model? tracing LLM lineage via spectral fingerprints in weight space. In *Conference on Language Modeling (COLM)*, 2026.

Pengzhou Cheng, Zongru Wu, Tianjie Ju, Wei Du, Zhuosheng Zhang, and Gongshen Liu. Transferring backdoors between large language models by knowledge distillation. *arXiv preprint arXiv:2408.09878*, 2024.

Xueqi Cheng and Yushun Dong. LatentRouter: Can we choose the right multimodal model before seeing its answer? *arXiv preprint arXiv:2605.11301*, 2026.

Moming Duan, Mingzhe Du, Rui Zhao, Mengying Wang, Yinghui Wu, Nigel Shadbolt, and Bingsheng He. Position: Current model licensing practices are dragging us into a quagmire of legal noncompliance. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Zhiyuan Fu, Junfan Chen, Lan Zhang, Ting Yang, Jun Niu, Hongyu Sun, Ruidong Li, Peng Liu, Jice Wang, Fannv He, and Yuqing Zhang. FDLLM: A dedicated detector for black-box LLMs fingerprinting. In *2025 IEEE 24th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 1374–1379, 2025. doi: 10.1109/TrustCom66490.2025.00159.

Zhehao Huang, Baijiong Lin, Jingyuan Zhang, Jingying Wang, Yuhang Liu, Ning Lu, Tao Li, and Xiaolin Huang. VL-RouterBench: A benchmark for vision-language model routing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022. URL https://arxiv.org/abs/2201.12086.

Ping Li, Trevor J. Hastie, and Kenneth W. Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 287–296, 2006.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2304.08485.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, 2022.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics*, pp. 2263–2279, 2022.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 2207–2217, 2021.

Xin Mu, Yu Wang, Yehong Zhang, Jiaqi Zhang, Hui Wang, Yang Xiang, and Yue Yu. Model provenance via model DNA. In *ECAI 2024: 27th European Conference on Artificial Intelligence*, Frontiers in Artificial Intelligence and Applications. IOS Press, 2024. doi: 10.3233/FAIA240595.

Ivica Nikolic, Teodora Baluta, and Prateek Saxena. Model provenance testing for large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 38370–38397, 2025. doi: 10.52202/085713-1146.

Dario Pasquini, Evgenios M. Kornaropoulos, and Giuseppe Ateniese. LLMmap: Fingerprinting for large language models. In *34th USENIX Security Symposium*, pp. 299–318, 2025.

Xiaoqi Qiu, Hao Zeng, Zhiyu Hou, and Hongxin Wei. Provable model provenance set for large language models. *arXiv preprint arXiv:2602.00772*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pp. 629–645, 2022.

Aman Singh Thakur and Rayan Khoury. Training leaves traces: Centered residual signatures for language model lineage verification. *arXiv preprint arXiv:2608.14929*, 2026.

Yun-Yun Tsai, Jia Hao Liang, Chuan Guo, Junfeng Yang, and Laurens van der Maaten. RAFP: Identifying LLM lineages via rare-region fingerprints. *arXiv preprint arXiv:2505.12682*, 2025.

Sudong Wang, Weiquan Huang, Xiaomin Yu, Zuhao Yang, Hehai Lin, Keming Wu, Chaojun Xiao, Chen Chen, Wenxuan Wang, Beier Zhu, Yunjian Zhang, and Chengwei Qin. Beyond SFT-to-RL: Pre-alignment via black-box on-policy distillation for multimodal RL, 2026. URL `https://arxiv.org/abs/2604.28123`.

Yubo Wang, Jianting Tang, Chaohu Liu, and Linli Xu. Tracking the copyright of large vision-language models through parameter learning adversarial images. In *International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=K7xpl3LZQp`.

Zhaomin Wu, Haodong Zhao, Ziyang Wang, Jizhou Guo, Qian Wang, and Bingsheng He. LLM DNA: Tracing model evolution via functional representations. In *International Conference on Learning Representations*, 2026.

Chengwei Xia, Fan Ma, Ruijie Quan, Yunqiu Xu, Kun Zhan, and Yi Yang. Echoes of ownership: Adversarial-guided dual injection for copyright protection in MLLMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Nicolas Yax, Pierre-Yves Oudeyer, and Stefano Palminteri. PhyloLM: Inferring the phylogeny of large language models and predicting their performances in benchmarks. In *International Conference on Learning Representations*, 2025.

Runpeng Yu and Xinchao Wang. Neural lineage. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 9556–9567, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952, 2023. doi: 10.1109/ICCV51070.2023.01100.

Jie Zhang, Dongrui Liu, Qian Chen, Linfeng Zhang, Yong Liu, Yu Qiao, and Jing Shao. REEF: Representation encoding fingerprints for large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SnDmPkOJ0T.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. doi: 10.1109/TPAMI.2024.3369699.

Yifei Zhao, Qian Lou, and Mengxin Zheng. SIF: Semantically in-distribution fingerprints for large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Xiaofan Zheng, Xinghao Wang, and Xiaojun Wan. Ghost in the shell: Synonym-aware logit shaping fingerprint for copyright protection of large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5538–5555, 2026a. doi: 10.18653/v1/2026.findings-acl.273.

Xiaofan Zheng, Xinghao Wang, and Xiaojun Wan. UMMF: Protecting copyright of large vision-language models through unlearning-based multimodal memorization fingerprint. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9497–9513, 2026b. doi: 10.18653/v1/2026.acl-long.429.

Sally Zhu, Ahmed M. Ahmed, Rohith Kuditipudi, and Percy Liang. Independence tests for language models. In *Forty-second International Conference on Machine Learning*, 2025.