Authenticating Matryoshka Nesting Dolls via MML-LLM-Zero-shot 3D Reconstruction

Yulia Kumar * and Srotriyo Sengupta

Posted Date: 2 February 2026

doi: 10.20944/preprints202602.0036.v1

Keywords: matryoshka nesting dolls; cultural heritage authentication; multimodal fusion; 20 vision–language models; Blum medial axis; zero-shot 3D reconstruction

 doi:10.20944/preprints202602.0036.v1

Article

Authenticating Matryoshka Nesting Dolls via MML-LLM-Zero-shot 3D Reconstruction

Yulia Kumar [1,2,*] and Srotriyo Sengupta [3]

[1] Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ
[2] Department of Computer Science and Technology, Kean University, Union, NJ
[3] Princeton University
* Correspondence: yulia.kumar@kean.edu

Abstract

This work presents a multimodal machine learning (MML) pipeline with zero-shot 3D completion for the digital preservation and authentication of Matryoshka nesting dolls (MND). A private collection is digitized as this novel multimodal dataset centered on turntable videos, augmented with single and group images and auxiliary physical and textual cues. A text modality is produced using Qwen3VL captions to enable video–text fusion and semantic motif analysis. A unimodal 2D baseline is established for fine-grained 8-way style recognition and a 3-way authenticity task, and is compared against multimodal configurations that incorporate learned text embeddings. To incorporate geometry as direct evidence, the pipeline integrates a silhouette-to-skeleton branch based on the Blum medial axis (BMA) and a convolutional autoencoder (CA) that reconstructs dense silhouettes from sparse skeletons, yielding a compact representation suitable for downstream 3D reasoning. The 3D pipeline is implemented along two complementary branches: zero-shot completion with a pretrained 3D prior (Hunyuan3D) and mesh-oriented skeletonization via a custom BMA procedure. Late fusion combines geometric and textual signals to improve decision confidence beyond appearance-only models. The framework supports authentication decisions with explicit geometric and semantic evidence and is transferable to other cultural artifacts. Potential applications include AR/VR, education, gaming, and assistive technologies. The code for this project is available upon request.

Keywords: matryoshka nesting dolls; cultural heritage authentication; multimodal fusion; vision–language models; Blum medial axis; zero-shot 3D reconstruction

1. Introduction

Expert appraisal of Matryoshka artifacts draws on regional schools, motifs, brushwork, wood and varnish finishing, and maker stamps. A comparatively underutilized signal is geometry: turned-wood profiles, lathe and knife marks, concentricity, and shell-fit tolerances that are most faithfully captured in 3D [1,5]. Controlled acquisition with photogrammetry and open pipelines such as COLMAP and MeshLab makes high-fidelity 3D feasible even for small glossy objects when capture is standardized [2–4]. In parallel, modern 2D encoders enable fine-grained recognition from appearance [6,7], while direct learning on point clouds and meshes enables shape-aware reasoning from 3D structure [8–10]. Vision–language models and OCR provide complementary textual cues from stamps, labels, and contextual descriptions [11–13]. For cultural heritage settings, reliable probabilities and interpretable evidence are essential; calibration and attribution methods offer a principled basis for risk-aware decisions and user-facing rationale [14–16]. The primary objective is preservation of a private collection by constructing its digital twin (Figure 1). The technical objective is to evaluate how 2D recognition, text grounding, and geometry-driven representations jointly support authenticity assessment. The study investigates whether modern 2D backbones can internalize authenticity cues from controlled turntable imagery, whether compact silhouette skeletons can serve as geometry-preserving compression, and

 doi:10.20944/preprints202602.0036.v1

whether pretrained 3D priors can complete missing geometry in a zero-shot regime. A complementary goal is to model nested structure by integrating geometry pipelines that support multi-shell reasoning and mesh refinement. The 3D reconstruction pipeline follows a standard SfM/MVS workflow: sparse reconstruction estimates camera poses and triangulates feature tracks via bundle adjustment, after which dense stereo computes depth and normal maps that are fused into a dense point cloud. This output supports downstream analysis, including zero-shot completion to infer occluded geometry without category-specific training. The remainder of the paper formalizes these components and evaluates their contributions in controlled experiments.

Figure 1. Snapshot of the Matryoshka dataset.

1.1. Research Questions

RQ1: Can a curated, turntable-based Matryoshka dataset support strong 2D multi-task baselines for 8-way style classification and 3-way authenticity prediction, and which backbones perform best on this benchmark?

RQ2: Can BMA-based skeletonization and a CA provide a compact geometric representation that faithfully reconstructs silhouettes and supports subsequent 3D reasoning?

RQ3: How do foundation models for vision and language perceive, describe, and generate Matryoshka artifacts, and how does LLM-generated text improve multimodal performance?

RQ4: How well does a zero-shot 3D completion prior trained on everyday objects transfer to axially symmetric cultural artifacts?

2. Related Work

Image-based 3D modeling is established for heritage documentation and small artifacts [1], with widely adopted SfM/MVS pipelines such as COLMAP [2,3] and mesh processing tools such as MeshLab [4]. Geometric analysis surveys in cultural heritage describe shape and curvature descriptors, acquisition constraints, and quality metrics aligned with lathe and shell-fit analysis [5]. For specular varnished surfaces, reflectance transformation imaging and polynomial texture mapping reveal micro-relief and toolmarks complementary to 3D meshes [17,18]. Transformer-based vision models and modernized convolutional neural networks (CNNs) are effective at fine-grained cues relevant to brushwork and motifs [6,7]. Vision–language pretraining provides semantic alignment that can be leveraged for description-driven retrieval and explanation [11]. Direct learning on point sets and meshes enables reasoning over local curvature and edge structure relevant to manufacturing signatures [8–10]. Transformer OCR and open pipelines address mixed scripts and scene text on curved

 doi:10.20944/preprints202602.0036.v1

or stamped surfaces [12,13]. Because neural networks can be miscalibrated, post-hoc temperature scaling is commonly used to improve reliability [14]. Attribution methods for images and point sets support user-facing rationale overlays aligned with cultural heritage requirements [15,16]. More detailed literature analysis and comparison are summarized in Table 1.

Table 1. Comparison with related works.

Ref.	What they did	Similarities	Differences
[36]	Medial-axis (MA) driven 3D shape segmentation using the MA transform to identify part junctions and produce weakly convex parts.	Uses skeleton-like structures for 3D shape analysis, aligned with BMA reasoning.	Geometry-only segmentation on generic shapes.
[37]	Deep segmentation of cultural heritage point clouds for architectural elements in HBIM workflows.	Uses 3D geometry and deep learning in a cultural heritage setting.	Targets large-scale architecture.
[38]	2D shape classification using skeleton path similarity and a Bayesian classifier.	Treats skeletons as robust shape descriptors for recognition.	Operates on generic 2D silhouettes.
[39]	Latent diffusion for 3D point clouds enabling high-quality generation and interpolation.	Motivates strong generative priors for 3D shapes.	Not for cultural heritage artifacts.
[40]	Textbook covering 3D acquisition, representation, registration, and cultural heritage digitization.	Provides foundational context for the 3D pipeline.	Survey and reference material, not a method.
[41]	Skeleton-based action recognition from joint trajectories.	Demonstrates skeletons as compact, invariant representations.	For temporal human actions.
[42]	Zero-shot 3D classification by adapting CLIP with realistic projections and LLM-generated prompts.	Aligns with the zero-shot 3D plus language axis.	Evaluated on generic benchmarks.
[43]	Zero-shot text-to-shape generation by coupling a shape autoencoder with CLIP-conditioned synthesis.	Connects text and 3D shape spaces without paired supervision.	Focuses on generic text-to-shape generation.

Lessons learned: The strongest precedent exists for individual modalities; the open methodological gap is an auditable fusion of appearance, text, and geometry for artifact authentication.

3. Methodology

An overview of the end-to-end pipeline is given in Figure 2. The workflow unifies three evidence streams: 2D appearance, text descriptions, and geometry. A 2D branch learns discriminative visual embeddings from turntable frames; a text branch encodes LLM-generated descriptions and recorded physical clues; and a geometry branch derives silhouettes and skeletons that support compact shape modeling and downstream 3D reasoning. Fusion modules combine modality-specific embeddings to produce final predictions for style and authenticity.

 doi:10.20944/preprints202602.0036.v1

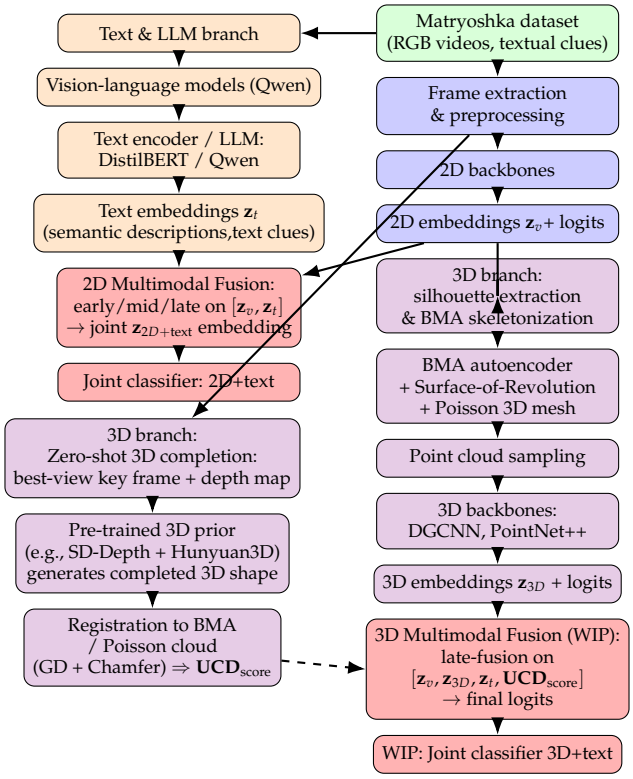


Figure 2. Methodology at-a-Glance.

3.1. Creating a Unique Dataset

The dataset is constructed from a private collection and is designed for controlled multi-view learning of Matryoshka nesting dolls (MND). The primary source is turntable video of individual artifacts; frames are extracted, quality-checked, and organized into class-labeled folders (Figure 1). Table 2 reports one representative configuration used for training and evaluation. Eight style classes capture the dominant visual regimes present in the collection, including artist-driven decorative variants, minimalist drafted patterns, merchandise-grade mass production, non-authentic replicas, non-Matryoshkas confounders, political portrait variants, religious iconography, and traditional Russian_Authentic schools. In addition, each sample is assigned a three-level authenticity label (RU, non-RU, unknown) to support both fine-grained and coarse-grained evaluation. Manual inspection records physical and textual authenticity cues, including country-of-origin markings and retail labels. These cues guide ambiguous cases and support text-conditioned modeling. A small manually labeled subset is also used to train a Teacher detector that enables self-corrected foreground extraction at scale, which stabilizes downstream BMA processing.

Table 2. Matryoshka dataset statistics: number of videos and labeled frames per class.

Class	# Videos	# Frames
artistic	22	4,087
drafted	21	4,622
merchandise	13	1,993
non_authentic	34	5,999
non_ Matryoshkas	13	1,940
political	5	782
religious	11	1,700
russian_authentic	49	6,264
Total	**168**	**27,387**

 doi:10.20944/preprints202602.0036.v1

Figure 3. Country-of-origin labels, retail stickers, and other physical clues used as signals for authenticity.

Lessons learned: Controlled acquisition and explicit capture of physical cues reduce label ambiguity; a small, high-quality manual subset is sufficient to bootstrap reliable large-scale self-correction.

3.2. LLM-Guided Matryoshka Generation

This study evaluates whether contemporary large language models (LLMs) and their associated image generators can synthesize Matryoshka imagery that is both visually plausible and class-consistent. To establish a controlled generative baseline, a two-stage prompting pipeline is adopted: a multimodal LLM produces structured, class-aware prompts, and a separate image model renders the corresponding images. In the reported configuration, Gemini-PRO generates prompts that are then executed by ChatGPT-4o, with one prompt template instantiated for each of the eight Matryoshka classes. This design improves reproducibility and makes it possible to audit how well the class taxonomy is internalized at the prompting stage. An example prompt used for the *Artistic* class is shown below.

> *"Please generate an image of a Matryoshka doll. Focus on an artistic type, emphasizing intricate, hand-painted details and unique thematic elements that showcase the craftsmanship beyond typical folk art. The doll should reflect a sophisticated design, potentially drawing from specific historical art movements or contemporary artistic interpretations while retaining the traditional nesting doll form."*

The prompt explicitly specifies stylistic constraints (detail level, ornament density, and optional art-historical influences) that act as high-level controls over palette, visual complexity, and mood. The template is further adjusted to produce coherent families of synthetic dolls per class, targeting consistency in pose, facial structure, and decorative grammar while preserving class-specific variability.

Figure 4. Synthetic examples for 8 classes: ChatGPT-4o (top), Gemini-2-PRO (mid), and Nano-Banana-PRO (bottom).

Zero-shot generation is also evaluated using the Google Nano-Banana-Pro model [35], issuing single-turn prompts without conversational context. Its outputs differ qualitatively from those produced by ChatGPT-4o and Gemini-PRO: in several cases, prompts intended for *non- Matryoshkas* produce plausible Matryoshka families, and political or religious themes are rendered in unpredictable ways. These failure modes indicate that current generators often preserve global "Matryoshka-ness" while weakening class boundaries. Despite these limitations, the results demonstrate substantial

 doi:10.20944/preprints202602.0036.v1

generative capacity and motivate future work on prompt alignment and controllable synthesis for cultural-heritage imagery.

Although synthetic generation is not the primary objective, it provides a controlled mechanism for distribution-shift experiments. A long-term direction is to blend real and synthetic Matryoshkas into a mixed dataset and evaluate how 2D, 3D, and multimodal backbones transfer to, or detect, synthetic counterparts. This enables stress-testing of authenticity cues and robustness under systematic perturbations in motif, style, and composition.

Lessons Learned: LLM-guided prompting enables reproducible, class-conditioned synthesis and supports rapid qualitative auditing across the full taxonomy. However, current generators frequently blur semantic boundaries between classes and do not reliably reproduce geometric and construction cues required for RU vs. non-RU authentication. Consequently, synthetic Matryoshka generation is treated primarily as a controlled stress-test source rather than a direct augmentation mechanism for authenticity modeling.

3.3. Applying Generative AI for Matryoshkas Generation

After several unsuccessful attempts with alternative generators and prompting strategies, SDXL–Juggernaut is adopted to produce Matryoshka images [34]. Figure 5 shows a representative grid for the *Artistic* category. The samples exhibit high photorealism, consistent lighting, and moderate intra-class diversity. Nevertheless, the outputs remain insufficient for correcting the class imbalance of the authentic dataset: pose and composition diversity are limited, subtle generation artifacts persist, and style averaging reduces the presence of distinctive workshop-specific features. Overall, the SDXL–Juggernaut pipeline, combined with grid visualization, enables rapid human-in-the-loop screening and clarifies where additional prompt constraints, guidance tuning, or targeted augmentation is required to better match the statistics of the real Matryoshka dataset.

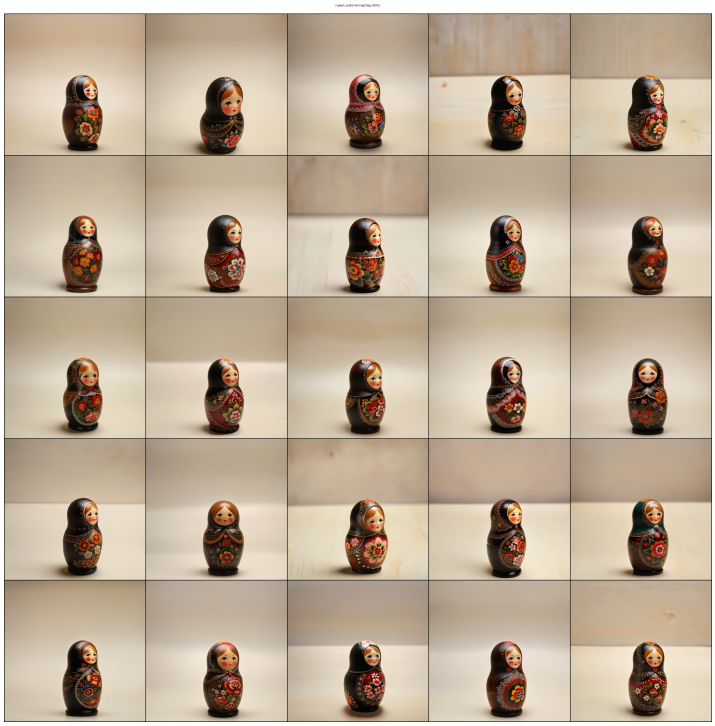


Figure 5. Matryoshkas generated via SDXL–Juggernaut.

Lessons Learned: SDXL–Juggernaut produces visually convincing Matryoshka imagery and supports efficient quality control via grid-based inspection, but it does not reliably generate the fine-grained construction and geometry signals needed for authenticity inference. In the current setting, the primary value is diagnostic: the synthetic grids expose which attributes are easily controlled (style, palette,

 doi:10.20944/preprints202602.0036.v1

motif density) and which remain underconstrained (pose variation, workshop-specific structure, and class-separating semantics).

3.4. 2D Pipeline for the Multi-Task Matryoshka Dataset

As a foundation for multimodal fusion and 3D geometry, we first establish strong *2D video-frame baselines*. Videos are stored in class-labeled folders and decoded at 5 fps. We remove near-duplicate frames via perceptual hashing and log per-frame quality-control (QC) statistics (brightness, blur via Laplacian variance, glare) into a metadata table. Figure 7 summarises global QC histograms, indicating that most frames are sufficiently well-exposed and sharp for recognition.

Each frame is annotated with *two labels*: (i) an 8-way stylistic class and (ii) an aggregated 3-way authenticity label (RU / non-RU / unknown). The 8-way distribution is imbalanced (Figure 6), but the authenticity task is closer to balanced. This multi-task design allows us to measure fine-grained style recognition while preserving the coarser authentication signal.

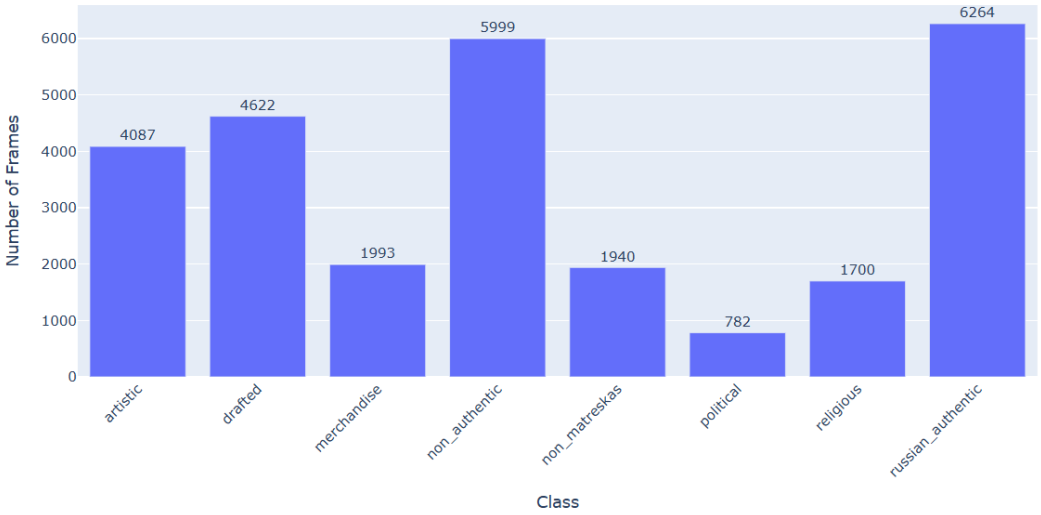


Figure 6. Number of extracted frames per semantic class.

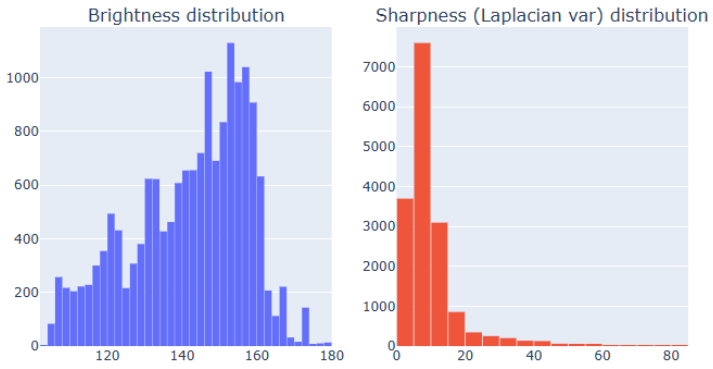


Figure 7. Quality-control histograms for all extracted frames: brightness (blue) and Laplacian variance (red).

Backbones and Heads

We evaluate five timm-pretrained 2D backbones: ConvNeXt-Tiny, VGG-16-BN, VGG-19-BN, Swin-Tiny, and ViT-Base [6,7,20–22]. Given an input frame $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$, the backbone produces an embedding $\mathbf{z} = f_b(\mathbf{x})$, followed by two linear heads:

$$\ell_{\text{cat}} = W_{\text{cat}}\mathbf{z} \in \mathbb{R}^8, \qquad \ell_{\text{auth}} = W_{\text{auth}}\mathbf{z} \in \mathbb{R}^3,$$

with softmax probabilities for the 8-way and 3-way tasks.

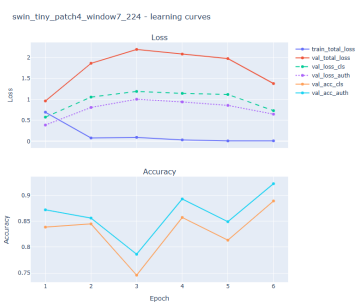
 doi:10.20944/preprints202602.0036.v1

Our training protocol is as follows. Frames are resized to 224×224, ImageNet-normalised, and augmented with mild flips, color jitter, and small affine transforms. We use a 70/15/15 train/val/test split, AdamW with cosine schedule (warm-up), initial learning rate 10^{-4}, weight decay 10^{-5}, and early stopping (patience $= 5$) on validation macro-AUPRC averaged across the two tasks. To mitigate 8-class imbalance, we use a *WeightedRandomSampler*. The objective is the sum of task-wise cross-entropy losses:

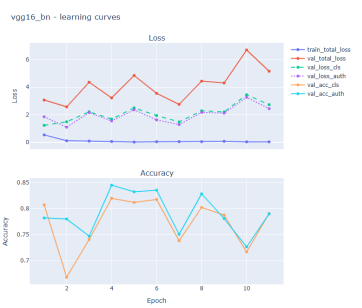
$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}}(\ell_{\mathrm{cat}}, y_{\mathrm{cat}}) + \mathcal{L}_{\mathrm{CE}}(\ell_{\mathrm{auth}}, y_{\mathrm{auth}}).$$

We report frame-level accuracy and macro-AUPRC for both tasks.

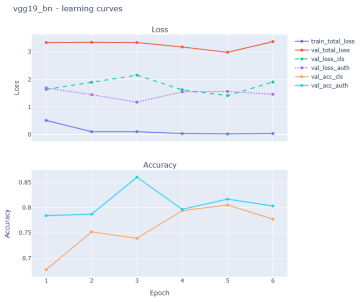
(**a**) ConvNeXt-Tiny (IN-22k).



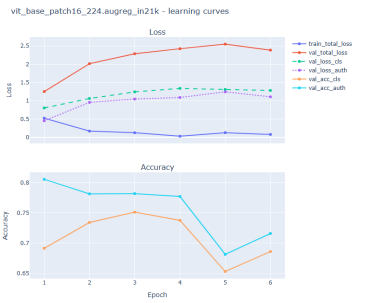
(**b**) Swin-Tiny.



(**c**) VGG16-BN.



(**d**) VGG19-BN.



(**e**) ViT-Base (AugReg IN-21k).

Figure 8. Loss and validation accuracy for 2D backbones: (**a**) ConvNeXt-Tiny (IN-22k); (**b**) Swin-Tiny; (**c**) VGG16-BN; (**d**) VGG19-BN; (**e**) ViT-Base (AugReg IN-21k).

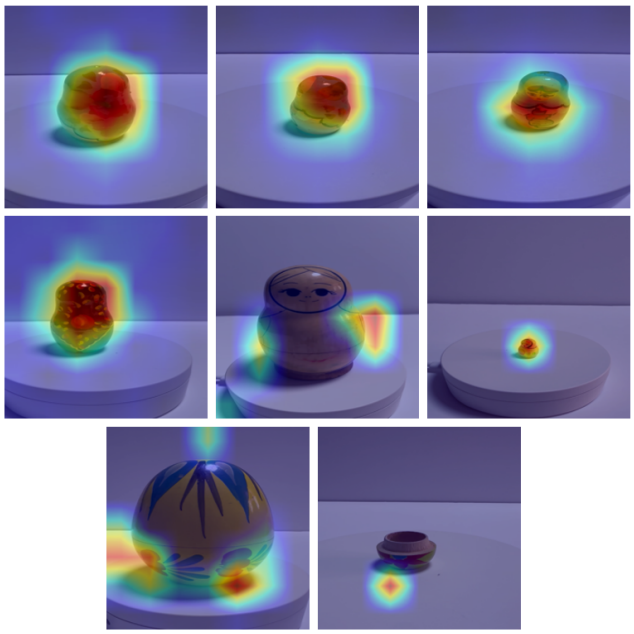


Figure 9. Grad-CAM visualization of the ConvNeXt-Tiny authenticity head.

 doi:10.20944/preprints202602.0036.v1

Figure 10 shows the validation confusion matrices for the 8-class and 3-way authenticity tasks across backbones. The main diagonal remains strong for most architectures, particularly ConvNeXt-Tiny and Swin-Tiny, while off-diagonal mass highlights common confusions such as *non_authentic* vs. *russian_authentic* and *artistic* vs. *russian_authentic*. For authenticity, the largest remaining error mode is *non-RU/replica* predicted as *RU*, which represents the most dangerous failure case for collectors and motivates the introduction of additional modalities (text and 3D).

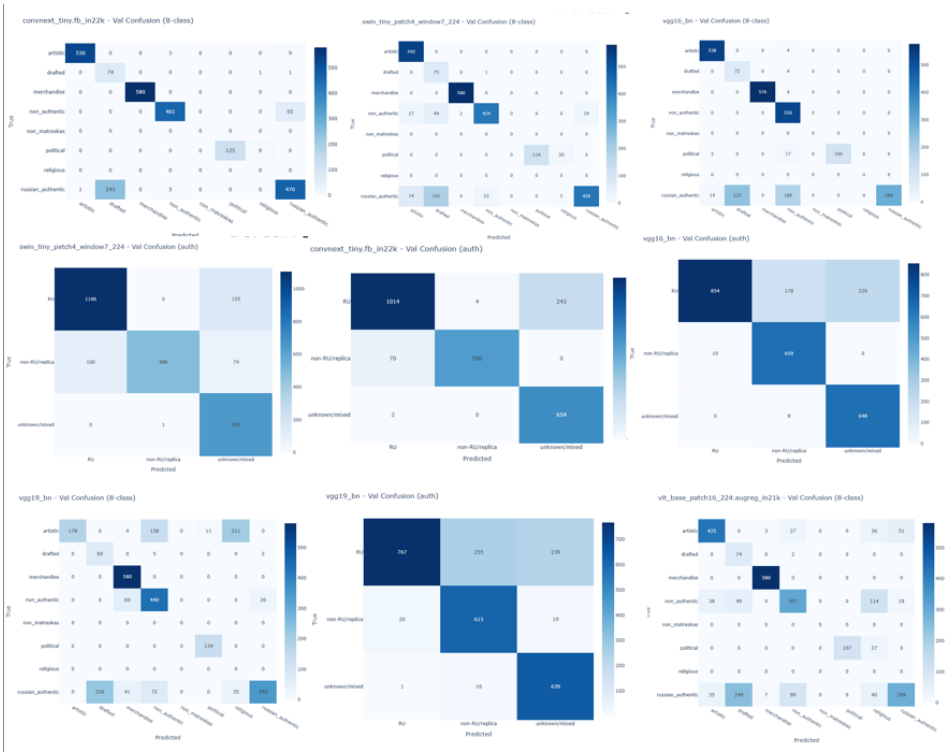


Figure 10. Validation confusion matrices for the 8-class category and the 3-way authenticity across 2D backbones.

To interpret the decision-making process of the ConvNeXt-based authentication model, we employ Gradient-weighted Class Activation Mapping (Grad-CAM, Figure 10). This technique visualizes the regions of the input image that most strongly influence the model's prediction for a specific class. Figure 9 presents a qualitative analysis of these activation maps across different categories. The visualisations reveal that the model consistently attends to the central features of the Matryoshka dolls, specifically the face and the central "apron" or belly region. This behaviour aligns with expert appraisal practices, as these areas typically contain the most distinctive stylistic motifs (e.g., *kokoshnik* headpiece or floral Khokhloma patterns) that define a specific regional school or artist. The model's ability to ignore background clutter suggests it has learned robust, object-centric features. When surface texture is minimal, such as with "drafted" (unpainted) dolls, the activation maps become more diffuse, adhering closely to the object's contours. This indicates a shift in the model's strategy from texture-based recognition to shape-based analysis when colour cues are absent. For non-Matryoshka objects like the egg in Figure 10, the model focuses on disparate surface textures and irregular geometries, correctly identifying them as out-of-distribution samples. Table 4 summarises validation and test performance for all five backbones on the multi-task Matryoshka 2D benchmark.

 doi:10.20944/preprints202602.0036.v1

Table 3. 2D backbone architectures used in the multi-task benchmark. "Feat. dim" denotes the penultimate embedding dimension fed to task heads.

Backbone	Feat. dim	Params	Notes
ConvNeXt-Tiny	768	27.8M	Hierarchical ConvNeXt blocks; strong accuracy/efficiency tradeoff.
Swin-Tiny	768	28.3M	Shifted-window attention; robust local-global features.
ViT-Base	768	86.6M	Global self-attention; strong capacity / heavier compute.
ResNet-50	2048	25.6M	Classic residual CNN.
EfficientNet-B0	1280	5.3M	Lightweight CNN.

Table 4. Performance of all backbones on the multi-task Matryoshka 2D benchmark.

Backbone	Val Acc		Val macro-AUPRC		Test Acc		Test macro-AUPRC	
	8-class	auth	8-class	auth	8-class	auth	8-class	auth
ConvNeXt-Tiny (IN-22k)	0.8692	0.8762	0.8628	0.9305	0.7778	0.7788	0.7824	0.6815
VGG-16 BN	0.8172	0.8351	0.8244	0.8797	0.6932	0.6908	0.5514	0.6267
VGG-19 BN	0.6775	0.7842	0.7817	0.8682	0.5304	0.6043	0.6440	0.5806
Swin-Tiny (patch4, window7, 224)	0.8386	0.8719	0.9670	0.9818	0.8298	0.8508	0.8823	0.8094
ViT-Base (patch16, 224, AugReg IN-21k)	0.6915	0.8056	0.9012	0.9446	0.7146	0.8284	0.7548	0.8015

ConvNeXt-Tiny and Swin-Tiny achieve the strongest overall performance, with Swin-Tiny yielding the highest test accuracy and macro-AUPRC across both tasks. ConvNeXt-Tiny provides a competitive convolutional baseline, while VGG-16 and VGG-19 underperform despite their larger parameter counts, suggesting that modern transformer-style architectures better capture the subtle stylistic variations in Matryoshka dolls. ViT-Base exhibits high macro-AUPRC, particularly on validation, but its test accuracy on the 8-class task remains below Swin-Tiny, likely reflecting differences in data efficiency and regularisation strategies.

Results and interpretability are discussed next. Table 4 summarises performance across backbones. We further analyse learning dynamics (Figure 8), confusion matrices (Figure 10), and Grad-CAM for the authenticity head (Figure 9) to verify that the model attends primarily to object-centric regions (face, apron/belly motifs) rather than background cues.

Lessons learned: Two conclusions are directly actionable for the rest of the paper.

(1) Modern backbones outperform legacy ConvNets. ConvNeXt-Tiny and Swin-Tiny consistently provide the strongest overall performance on both the 8-way style task and the 3-way authenticity task, while VGG-16/19 underperform despite higher parameter counts. This supports using modern ConvNet/Transformer families as the default 2D feature extractors for fusion.

(2) The dominant failure mode is authenticity ambiguity. Even with strong 2D performance, the most consequential errors remain *non-RU/replica predicted as RU*. High-quality replicas can match texture, palette, and brushwork, making RGB-only evidence insufficient for risk-sensitive authentication. This directly motivates adding complementary modalities: (i) text cues (labels, stamps, inscriptions, LLM-generated descriptions) and (ii) geometry (BMA skeletons, profiles, and 3D shape).

3.5. Improving Text Modality via LLMs

We construct a *text modality* by converting representative Matryoshka frames into short, structured captions that can be used as semantic inputs for text-only and multimodal models. Early experiments relied on manual prompting of proprietary foundation models (Figure 11), which was useful for prompt design but not scalable. To caption the full dataset without paid APIs, we switch to open-source vision–language models from the Qwen family. For each video folder, we select a representative frame and query a Qwen-VL model with a fixed captioning prompt. We first tested Qwen2-VL-2B-Instruct and then upgraded to Qwen3-VL-8B-Instruct, which consistently produced more specific descriptions (e.g., materials, finishes, construction cues) and more explicit statements about authenticity.

 doi:10.20944/preprints202602.0036.v1

Figure 11. Manual prompting of Gemini-Pro for captions.

Captioning prompt *"Please provide a concise, two-sentence description of the Matryoshka doll (or doll set) in this image. State its place in the set (e.g., smallest, middle, outer shell, full family). Focus on visual details, style, and notable features that might indicate region or school of craftsmanship. Keep the description around 200–300 characters. Also comment if it appears authentic Russian or not."*

The generated captions are stored alongside video identifiers and used as "image reports." In this subsection, we evaluate a *text-only* pipeline (no pixels) to isolate the semantic signal contained in captions before introducing multimodal fusion.

In our study, we compare the quality of Qwen2-VL vs. Qwen3-VL in creating accurate and usefult descriptions. We noticed that Qwen2-VL often produces generic descriptions (colors, patterns, cultural usage) and tends to implicitly assume "traditional Russian nesting doll" semantics. In contrast, Qwen3-VL more frequently describes concrete physical details and flags modern construction cues that are relevant to RU vs. non-RU labels (Table 5).

Table 5. Qwen2-VL vs. Qwen3-VL comparison for the same Matryoshka.

Aspect	Qwen2-VL	Qwen3-VL
Caption	The Matryoshka doll in the image is a vibrant red and blue set, featuring intricate floral and paisley patterns. It is a traditional Russian nesting doll, known for its ability to be stacked and opened to reveal smaller versions of itself. The doll is designed with a playful and whimsical aesthetic, often used in Russian folk art and traditional Russian culture. It is a collectible item, often given as a gift or used in celebrations like New Year's Eve.	This is the outer shell of a traditional Russian Matryoshka doll, featuring rich burgundy enamel with hand-painted red roses, green leaves, and gold scrollwork against a dark blue lower band. Its ornate floral design and glossy finish reflect classic Russian folk art, though the base's modern white plastic suggests a contemporary or mass-produced version rather than an authentic antique.
Detail level	General colors/patterns and broad cultural framing.	Specific colors, materials, finishes, and construction cues.
Auth. cues	Implicitly assumes "traditional Russian"; rarely qualifies authenticity.	Explicitly contrasts folk-art styling with modern manufacturing indicators.

We train a shared text encoder with two heads (8-way style and 3-way authenticity) using only Qwen3 captions (column *caption_qwen3*). We compare four encoder families: a Qwen embedding model, a compact sentence-transformer, a recent LLM encoder (Phi-3), and DistilBERT to maintain

 doi:10.20944/preprints202602.0036.v1

comparability with our initial fusion baseline. We use short inputs (96 tokens) and differential learning rates (lower for the encoder, higher for the heads) to stabilise fine-tuning on a small dataset (Table 6).

Table 6. Text-only configuration (Qwen3 captions).

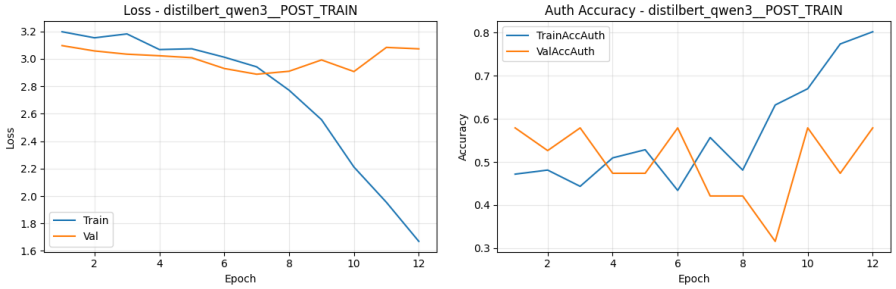
Parameter	Value	Details
Backbones Tested	4	Qwen3-Embedding-0.6B, all-MiniLM-L6-v2, Phi-3-mini-4k-instruct, DistilBERT-uncased
Input Text Source	*caption_qwen3*	Qwen3-VL-8B captions
Input Size	96 tokens	*MAX_TEXT_LEN*
Batch Size	4	Memory-controlled
Learning Rate Scheme	Differential LR	Encoder (5×10^{-6}), heads (2×10^{-4})
Dropout	0.3	On shared text representation
Loss	CE (8-class + auth)	Class-weighted for imbalance

Table 7 reports test macro-F1. MiniLM achieves the strongest overall performance, while DistilBERT remains substantially weaker even after optimization. Across models, the main failure mode is *non-RU misclassified as RU*, which is consistent with a captioning bias: decorative replicas are frequently described using "traditional Russian" cues unless the caption explicitly flags modern manufacturing details.

Table 7. Text-only results on Qwen3 captions (test macro-F1).

Text Encoder	8-class	auth	Model Status
Qwen/Qwen3-Embedding-0.6B	0.26	0.58	Optimized (CPU)
all-MiniLM-L6-v2	0.43	0.60	Pilot (pre-optimized LR)
Phi-3-mini-4k-instruct	0.08	0.43	Optimized (unstable)
DistilBERT-uncased	**0.31**	**0.39**	Optimized (CPU)

For DistilBERT, training stabilises early and then overfits: validation loss reaches a minimum and subsequently increases, while authenticity accuracy fluctuates (Figure 12). The corresponding confusion matrices (Figure 13) show that many true non-RU samples are predicted as RU, reflecting the semantic prior embedded in captions.



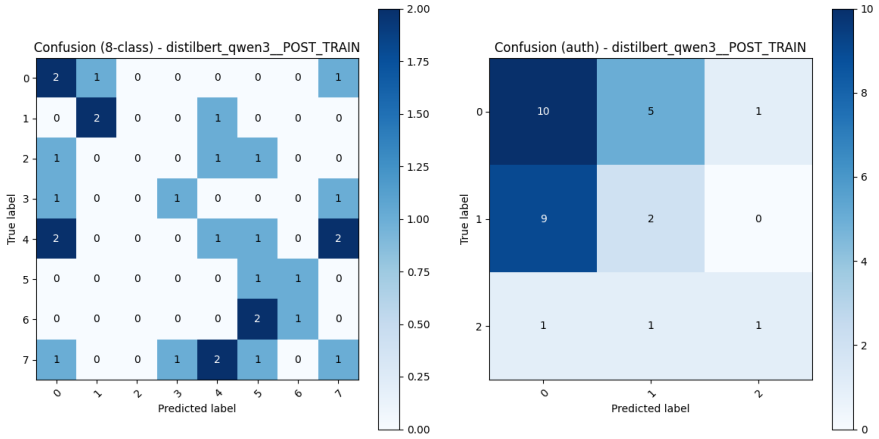
(**a**) Loss curves. (**b**) Authenticity accuracy.

Figure 12. Training statistics for DistilBERT on Qwen3 captions: (**a**) loss curves; (**b**) authenticity accuracy.

Lessons Learned: (1) Caption quality matters, but it does not remove semantic priors. Qwen3-VL substantially improves descriptive specificity relative to Qwen2-VL, yet captions still often default to "Russian" framing for visually plausible replicas unless modern construction cues are explicitly mentioned. (2) Encoder choice dominates text-only performance on small data. Compact sentence-transformers (MiniLM) are more data-efficient and stable than larger instruction-tuned LLM encoders

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 February 2026 doi:10.20944/preprints202602.0036.v1

in this setting; DistilBERT remains a useful reference baseline but is not competitive for authenticity. (3) Text alone is insufficient for verification-grade authenticity. Because captions can encode stylistic assumptions and miss subtle physical evidence, text-only models exhibit dangerous errors (non-RU → RU). This motivates multimodal fusion, where image/geometry features can override caption priors.



(**a**) 8-class confusion matrix (test). (**b**) Authenticity confusion matrix (test).

Figure 13. Confusion matrices for DistilBERT on Qwen3 captions (test): 8-class; authenticity.

Our motivation for Multimodal Fusion is a follows. Text-only authenticity performance remains well below the multimodal results (Section 3.6), indicating that captions provide useful context but cannot reliably resolve RU vs. non-RU in isolation. We therefore use text as a complementary modality rather than a primary authenticity signal.

3.6. Multimodal Fusion of Video Features and Text

We now analyse the behaviour of the *multimodal* Matryoshka classifier, in which video frames and Qwen-generated captions are fused to jointly predict (i) the 8-way semantic category and (ii) the 3-way authenticity label (RU / non-RU / unknown). This fusion stage is primarily motivated by the limitations of text-only inference: captions can be stylistically rich yet semantically biased, and in particular may over-ascribe "Russian" attributes to decorative replicas. Multimodal fusion mitigates this failure mode by forcing consistency between visual evidence and language cues.

All multimodal experiments follow a common pipeline: (i) sampling $T = 8$ RGB frames per video at 224×224 resolution, (ii) a *timm* 2D backbone for the video stream and DistilBERT for the caption stream, (iii) early, mid, or late fusion of projected video and text features, (iv) weighted cross-entropy losses for the 8-way class and 3-way authenticity labels, and (v) a 70/15/15 stratified split over the eight semantic classes. Training uses AdamW with a cosine schedule (with warmup), ImageNet normalization, strong color + geometric augmentation, and early stopping on mean macro-AUPRC (patience $= 5$). Here we report three representative 2D backbones: *convnext_tiny.fb_in22k*, *vgg16_bn*, and *vgg19_bn*.

For each Matryoshka video, the visual backbone produces a pooled video embedding h_v, while the caption stream yields a text embedding h_t from the generated description. We project both embeddings into a shared d_{fuse}-dimensional space using linear layers $W_v, W_t \in \mathbb{R}^{d_{\text{fuse}} \times d_{\text{in}}}$:

$$z_v = W_v h_v, \qquad z_t = W_t h_t.$$

On top of these projected features, we implement three fusion strategies.

For early fusion (MLP-based) the projected features are concatenated and passed through a multi-layer perceptron (MLP) ϕ:

$$z_{\text{joint}} = \phi\big(\text{Concat}(z_v, z_t)\big). \tag{1}$$

 doi:10.20944/preprints202602.0036.v1

The joint representation z_{joint} is then fed to a classifier for both tasks.

For Mid Fusion We treat z_v and z_t as a short token sequence $Z_{\text{seq}} = [z_v, z_t]$. A Transformer encoder layer $\mathcal{T}$ applies self-attention across modalities, followed by mean pooling:

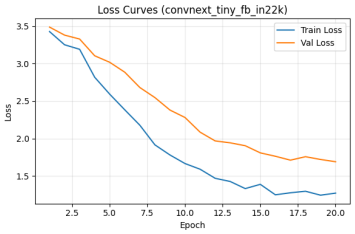
$$z_{\text{joint}} = \text{Mean}\big(\mathcal{T}(Z_{\text{seq}})\big). \tag{2}$$

For Late Fusion, we consider logit-level fusion where no explicit joint embedding is formed. Separate classification heads are applied to h_v and h_t, and logits are averaged:

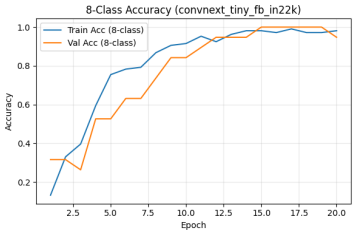
$$y_{\text{pred}} = \tfrac{1}{2}\Big(\text{Head}_v(h_v) + \text{Head}_t(h_t)\Big). \tag{3}$$

This setup keeps vision and language pipelines loosely coupled while allowing modality-specialized heads.

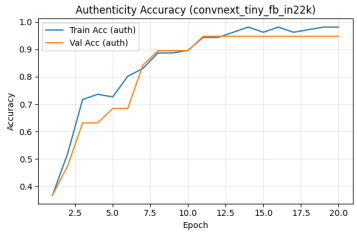
Figure 14 summarizes the optimization behaviour for ConvNeXt-Tiny and the VGG variants in the multimodal setting. ConvNeXt-Tiny exhibits smooth and stable optimization: both training and validation loss decrease without divergence, while 8-class and authenticity accuracy rise quickly and remain tightly coupled between train/val curves. VGG16-BN converges reliably but shows mild late-epoch oscillations in the 8-class task, consistent with small-data sensitivity. VGG19-BN reduces training loss while validation loss plateaus earlier, suggesting mild over-capacity under limited supervision; nonetheless, authenticity remains comparatively easier to fit and stays high across splits.



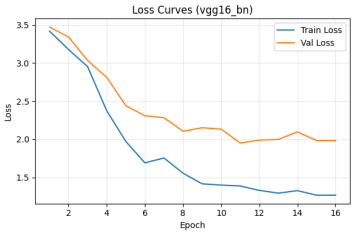
(**a**) Loss (ConvNeXt-T).



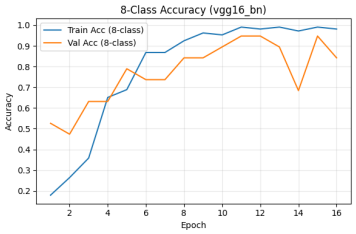
(**b**) 8-class accuracy (ConvNeXt-T).



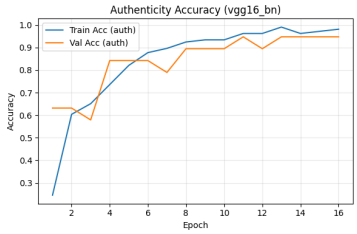
(**c**) Authenticity accuracy (ConvNeXt-T).



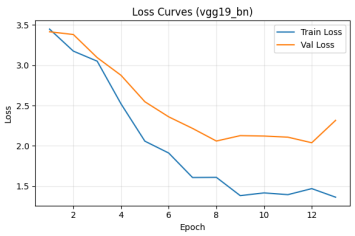
(**d**) Loss (VGG16-BN).



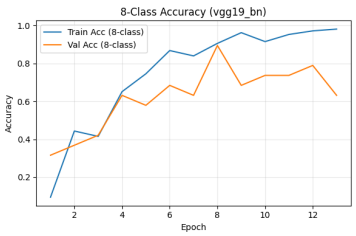
(**e**) 8-class accuracy (VGG16-BN).



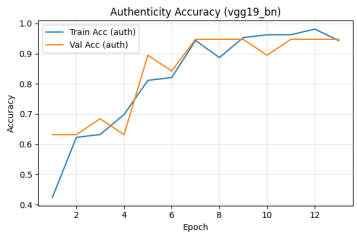
(**f**) Authenticity accuracy (VGG16-BN).



(**g**) Loss (VGG19-BN).



(**h**) 8-class accuracy (VGG19-BN).

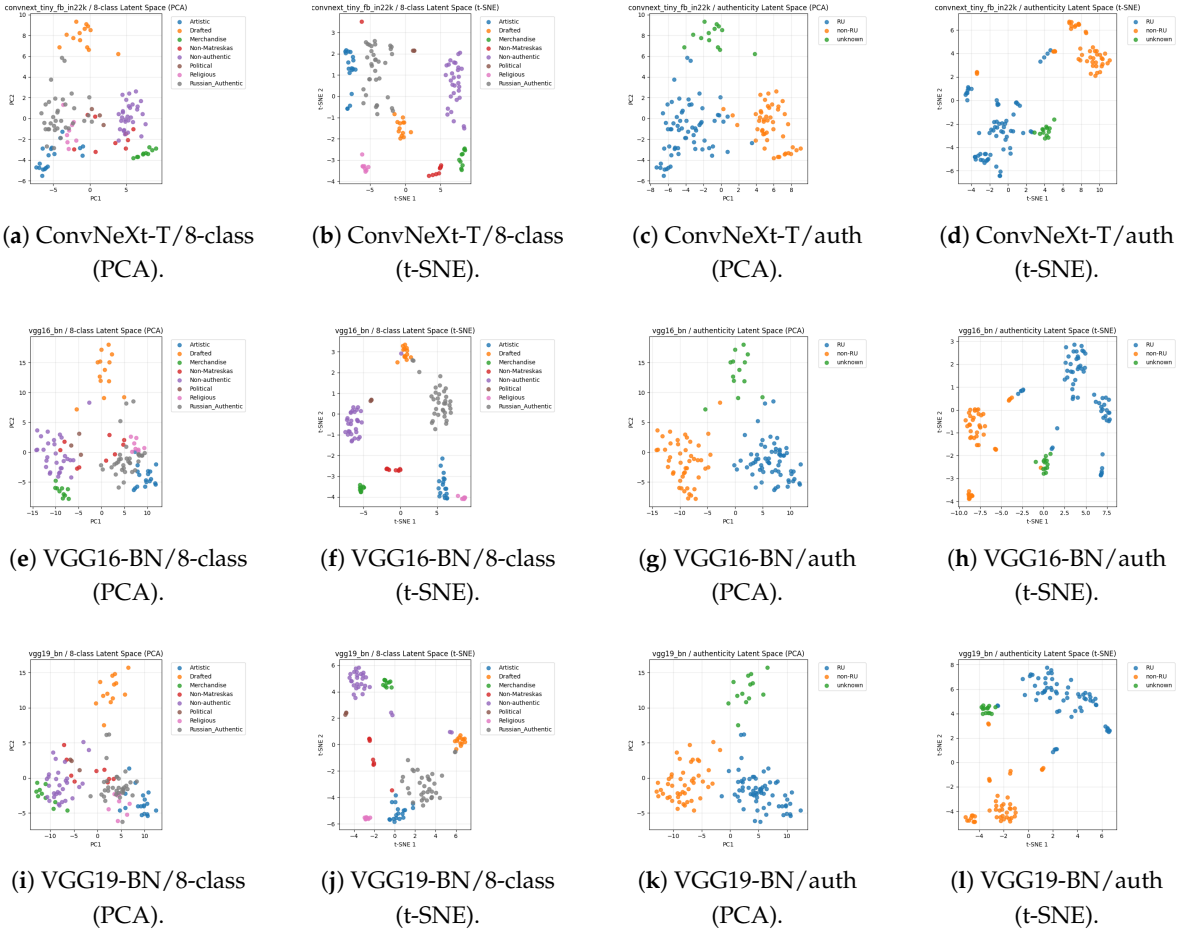


(**i**) Authenticity accuracy (VGG19-BN).

Figure 14. Training dynamics for three multimodal backbones: ConvNeXt-Tiny (top row), VGG16-BN (middle), and VGG19-BN (bottom). Columns show loss, 8-class accuracy, and authenticity accuracy.

 doi:10.20944/preprints202602.0036.v1

To better understand what the multimodal networks learn, we visualise fused embeddings (video + text) using PCA and t-SNE. ConvNeXt-Tiny yields compact, well-separated clusters for many semantic classes, while authenticity embeddings typically partition into three regions corresponding to RU, non-RU, and unknown. VGG16-BN remains interpretable but shows slightly increased overlap for visually adjacent categories (e.g., Political vs. Religious). VGG19-BN clusters are still meaningful but manifolds tend to lie closer together, consistent with its higher validation loss under small-data supervision.



(**a**) ConvNeXt-T/8-class (PCA). (**b**) ConvNeXt-T/8-class (t-SNE). (**c**) ConvNeXt-T/auth (PCA). (**d**) ConvNeXt-T/auth (t-SNE).

(**e**) VGG16-BN/8-class (PCA). (**f**) VGG16-BN/8-class (t-SNE). (**g**) VGG16-BN/auth (PCA). (**h**) VGG16-BN/auth (t-SNE).

(**i**) VGG19-BN/8-class (PCA). (**j**) VGG19-BN/8-class (t-SNE). (**k**) VGG19-BN/auth (PCA). (**l**) VGG19-BN/auth (t-SNE).

Figure 15. Fused video+text latent spaces: PCA vs. t-SNE across backbones. (**a–d**) ConvNeXt-Tiny, (**e–h**) VGG16-BN, (**i–l**) VGG19-BN.

We employ a weighted cross-entropy loss to address class imbalance. The total objective $\mathcal{L}_{\text{Total}}$ is the sum of the weighted cross-entropy losses for the 8-class and authenticity tasks:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{CE}}(\mathbf{p}_8, \mathbf{y}_8) + \mathcal{L}_{\text{CE}}(\mathbf{p}_{\text{Auth}}, \mathbf{y}_{\text{Auth}}),$$

where the class-weighted cross-entropy is

$$\mathcal{L}_{\text{CE}}(\mathbf{p}, \mathbf{y}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathcal{W}_k \cdot y_{i,k} \cdot \log(p_{i,k}),$$

and $\mathcal{W}_k$ represents the inverse-frequency weight for class k.

Early fusion is sensitive to cross-modal alignment and feature scaling. Mid fusion enables explicit cross-modal interaction via self-attention. Late fusion averages logits from separate heads, yielding robustness and a modular design.

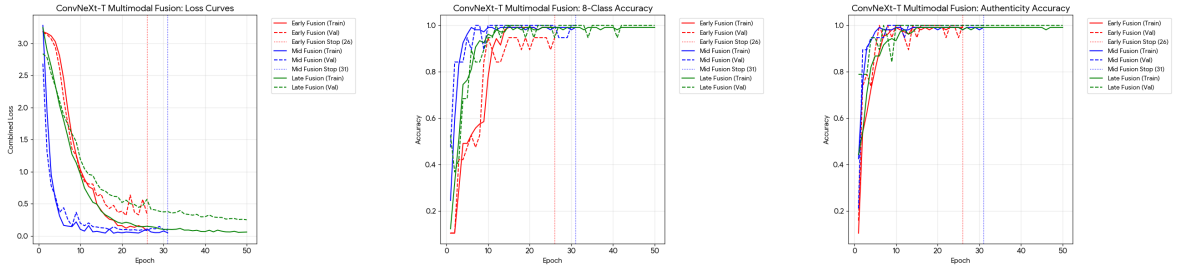
 doi:10.20944/preprints202602.0036.v1

Training uses early stopping (patience = 5 epochs) based on validation loss. As expected, Mid and Early fusion models stop earlier, while the Late model uses the full 50-epoch budget.

Table 8. ConvNeXt-T fusion results (multimodal).

Method	Stop Epoch	Best ValLoss	Best $\overline{\text{Acc}}_{\text{val}}$	Test Acc (8)	Test Acc (Auth)
Early	**26**	0.3169	0.974	0.867	0.933
Mid	31	**0.0815**	**1.000**	**0.900**	0.900
Late	50	0.2489	**1.000**	**0.900**	**0.967**

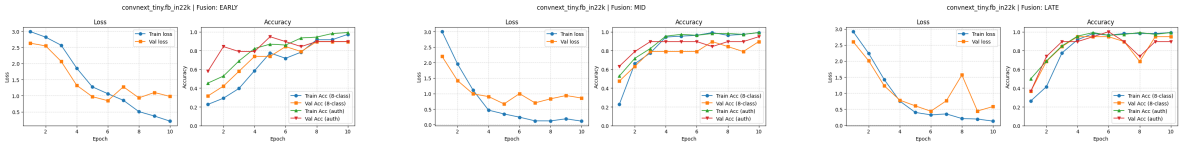
Figure 16 compares loss and accuracy curves across fusion strategies. Mid fusion reaches the lowest validation loss and matches the best validation accuracy, indicating highly efficient cross-modal interaction. Early fusion converges quickly but stalls and shows mild overfitting. Late fusion exhibits the most stable training and achieves the best test authenticity accuracy. In practice, we treat late fusion as the default multimodal configuration due to robustness and modularity.



(**a**) Loss curves: train vs. val. (**b**) 8-class accuracy. (**c**) Authenticity accuracy.

Figure 16. Training statistics for ConvNeXt-T fusion: (**a**) loss curves; (**b**) 8-class accuracy; (**c**) authenticity accuracy.

For completeness, Figure 17 provides the per-strategy training/validation summaries (Early/Mid/Late).



(**a**) Early fusion loss and accuracy. (**b**) Mid fusion loss and accuracy. (**c**) Late fusion loss and accuracy.

Figure 17. Training and validation loss/accuracy for three fusion strategies combining ConvNeXt-Tiny video features with Qwen3-VL text embeddings: (**a**) early fusion; (**b**) mid fusion; (**c**) late fusion.

To interpret which visual regions most influence authenticity decisions, we employ Grad-CAM on the ConvNeXt-based authenticity head. Qualitatively, activation maps emphasize the Matryoshka's face and central apron/belly motifs (areas that carry distinctive stylistic signatures), become more diffuse and contour-following for minimally textured "drafted" dolls, and attend to irregular textures/geometry for non-Matryoshka objects.

 doi:10.20944/preprints202602.0036.v1

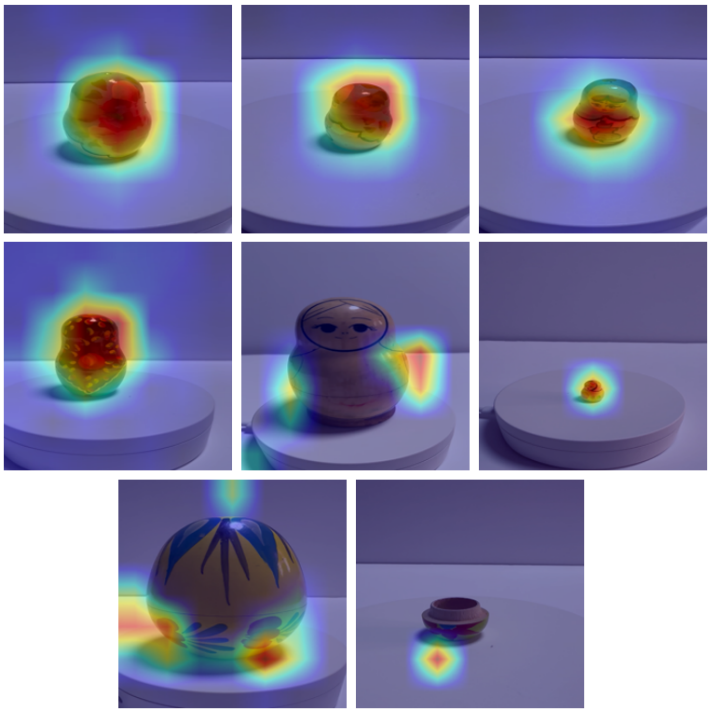


Figure 18. Grad-CAM visualisation for ConvNeXt-Tiny.

Overall, the multimodal fusion pipeline achieves high accuracy on both tasks, learns well-structured fused latent spaces aligned with semantic and authenticity labels, and exhibits stable optimization behaviour. Late fusion consistently yields strong validation performance, making it our preferred multimodal architecture. These results, together with the unimodal baselines from Section 3.4 and the text-only analysis in Section 3.5, motivate ConvNeXt-Tiny (and Swin-Tiny) as strong starting points for subsequent experiments that incorporate 3D geometry and skeleton-based descriptors.

3.7. Shape Analysis — Blum Medial Axis (BMA), YOLO Cropping, and Skeleton-Based Compression

In addition to RGB-based classification, we introduce an explicit *shape modality* by extracting Blum Medial Axis (BMA) skeletons from binary silhouettes of Matryoshka frames [24–26]. The motivation is two-fold: (i) to obtain a compact geometric descriptor that reduces reliance on texture cues (critical for authenticity), and (ii) to enable a high-ratio geometric compression pipeline in which sparse skeletons can be mapped back to dense silhouettes as a precursor to 3D reconstruction.

To create high-quality skeletons We first localize the Matryoshka in each frame using an Ultralytics YOLO detector, then crop to the predicted ROI prior to binarization and skeletonization. This isolates the object, stabilizes silhouette extraction, and reduces spurious skeleton branches caused by background edges. We use a *teacher-assisted dataset build* strategy: a pretrained teacher detector generates bounding boxes over the frame pool, from which we subsample a curated set for the downstream BMA and reconstruction experiments. A representative grid of YOLO ROI predictions is shown in Figure 19.

Because ROI errors can propagate into the silhouette/skeleton pipeline, we additionally perform *manual QA and correction* for selected samples using the `makesense.ai` annotation tool. Figure 20 shows the labeling interface and an example manual ROI box.

After ROI cropping, each frame is binarized to obtain a foreground mask, the dominant connected component is retained, the boundary is extracted, and the Blum Medial Axis is computed. Figure 21 shows a representative skeleton grid across categories using the ROI-cropped pipeline.

 doi:10.20944/preprints202602.0036.v1

Figure 19. Teacher-assisted ROI localization (YOLO).

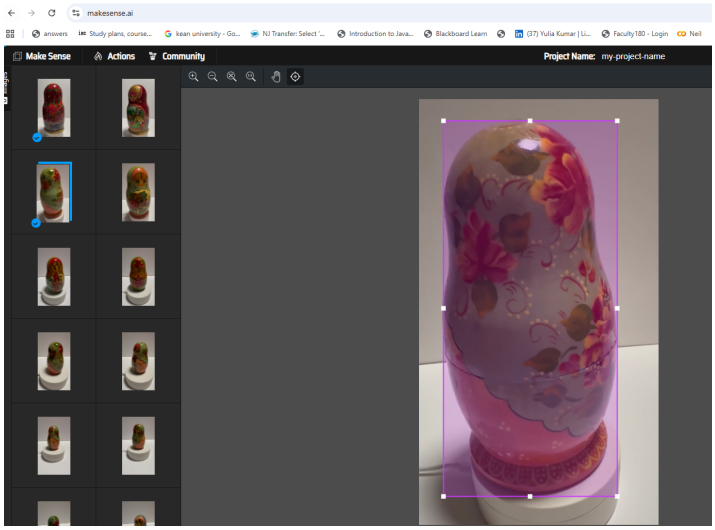


Figure 20. Manual labeling via The `makesense.ai`

In addition to the medial axis, we derive a compact *radius/shape profile* from the boundary geometry. The resulting profiles (Figure 22) capture global curvature and waist/bulge structure, offering a lightweight 1D descriptor that complements the skeleton representation.

 doi:10.20944/preprints202602.0036.v1

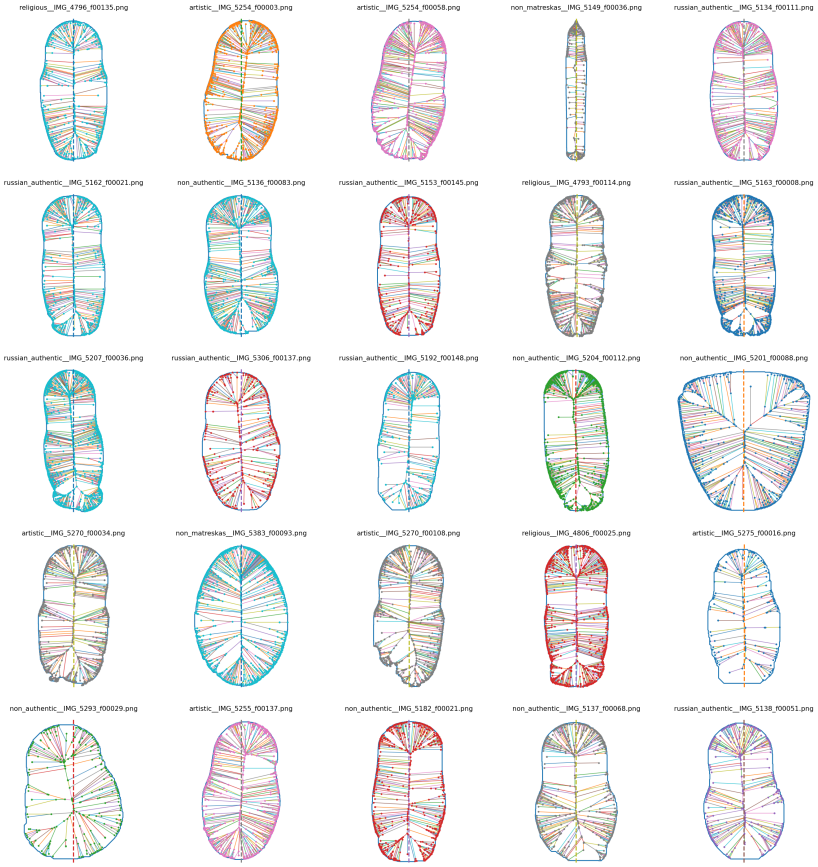


Figure 21. BMA skeleton grids (ROI-cropped).

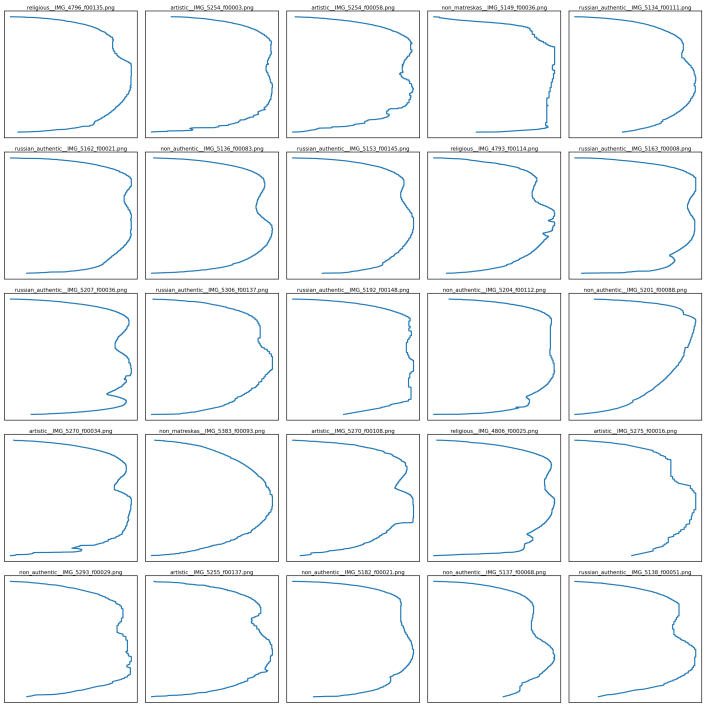


Figure 22. Boundary/radius-profile grids. capturing characteristic shape variations (e.g., base expansion and shoulder curvature) across classes.

For Details of skeleton creation see Algorithm 1.

 doi:10.20944/preprints202602.0036.v1

Algorithm 1 YOLO-ROI BMA Skeleton Extraction and Skeleton-to-Silhouette Compression

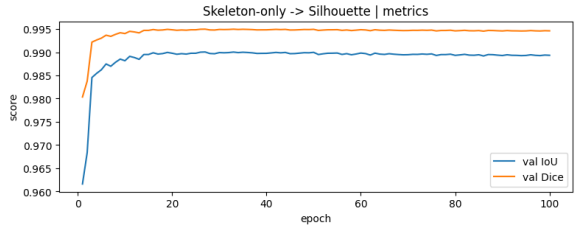
Require: Frame pool or videos root $\mathcal{V}$; YOLO teacher weights θ_{yolo}; ROI crop policy (padding p); binarization + cleanup ops; max samples N; autoencoder $\mathcal{A}_\theta$; training hyperparameters.
Ensure: Dataset index `index.csv`; best checkpoints for skeleton-only and overlay modes.

Mount Drive and configure Ultralytics settings (`v0.0.6`).
Load YOLO teacher: $\theta_{yolo} \leftarrow$ `best.pt`.
Initialize empty dataset list $\mathcal{D} \leftarrow \emptyset$.
for each frame I in candidate pool **do**
    Predict ROI box $b \leftarrow YOLO(I; \theta_{yolo})$.
    if manual label exists for frame I **then**
        Override $b \leftarrow b_{\text{manual}}$ (from `makesense.ai` export).
    end if
    Crop ROI: $I_{roi} \leftarrow$ `crop`(I, b, p).
    Build binary mask $M \leftarrow$ `binarize`(I_{roi}); keep largest component; optional morph. cleanup.
    Extract boundary ∂M and compute BMA skeleton $S \leftarrow$ `BMA`(∂M).
    Rasterize skeleton image $X_{\text{skel}} \leftarrow$ `rasterize`(S).
    Build overlay input $X_{\text{ov}} \leftarrow$ `overlay`$(I_{roi}, X_{\text{skel}})$.
    Append record $(X_{\text{skel}}, X_{\text{ov}}, M, \text{meta})$ to $\mathcal{D}$.
    if $|\mathcal{D}| = N$ **then break**
    end if
end for
Save dataset index to `index.csv`.
Split $\mathcal{D}$ into train/val; create loaders.
Train mode 1 (skeleton-only): input X_{skel} (1 channel), target M.
Initialize $\mathcal{A}_\theta$; optimize BCE (or BCE+Dice) to reconstruct M.
for epoch $= 1$ to E **do**
    Train on batches; evaluate val IoU/Dice; save best`ckpt_skeleton_best.pt`.
end for
Train mode 2 (overlay-crop): input X_{ov} (3 channels), target M.
Reinitialize or finetune $\mathcal{A}_\theta$; same objective.
for epoch $= 1$ to E **do**
    Train on batches; evaluate val IoU/Dice; save best
    `ckpt_overlay_best.pt`.
end for
Return best checkpoints and index.

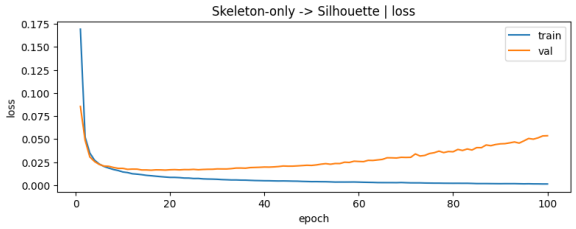
3.8. Skeleton → Silhouette Reconstruction as Geometric Compression

To validate that the Blum Medial Axis (BMA) skeleton preserves sufficient semantic geometry for downstream modeling, we train a lightweight convolutional autoencoder to reconstruct the dense binary silhouette from sparse BMA-derived inputs. This experiment treats the skeleton as an explicit geometric compression medium: although it drastically reduces pixel density, it preserves connectivity, symmetry, and global shape cues that should be recoverable by a learnable decoder. We evaluate two reconstruction regimes: (1) **Skeleton-only → silhouette**: a single-channel skeleton raster as input. (2)) **Overlay-crop RGB → silhouette**: a three-channel ROI crop of the doll with the skeleton overlaid. Figures 23 and 24 report convergence and validation overlap (IoU/Dice) for both regimes using the exact run outputs produced by our pipeline. The skeleton-only regime converges rapidly and reaches near-perfect overlap, indicating that the mapping from medial-axis topology to dense silhouette is stable and highly learnable for Matryoshka geometries. Validation IoU/Dice saturate early and the loss decreases smoothly, reflecting high geometric recoverability from sparse BMA inputs. The ROI-cropped overlay regime also achieves strong overlap, but converges to slightly lower IoU/Dice, consistent with appearance variability and residual texture/illumination effects present in RGB overlays.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 February 2026 doi:10.20944/preprints202602.0036.v1

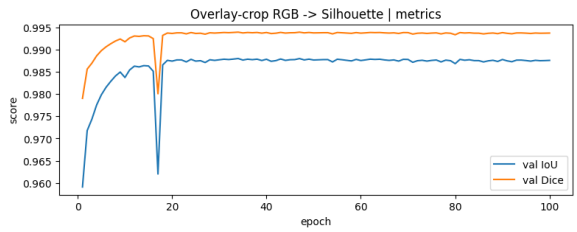


(**a**) Skeleton-only → silhouette: metrics.

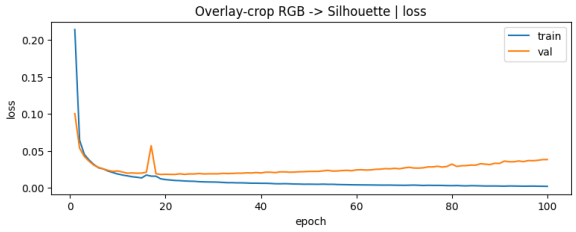


(**b**) Skeleton-only → silhouette: loss.

Figure 23. Skeleton-only reconstruction: (**a**) metrics; (**b**) loss.



(**a**) Overlay-crop RGB → silhouette: metrics.



(**b**) Overlay-crop RGB → silhouette: loss.

Figure 24. Overlay-crop reconstruction: (**a**) metrics; (**b**) loss.

Qualitative reconstructions are shown in Figure 25. The autoencoder successfully "inflates" thin skeleton traces into coherent silhouettes, recovering the bulbous body, shoulder transitions, and base geometry. These results support the use of skeleton latents as a compact geometric modality suitable for downstream fusion and 3D integration.

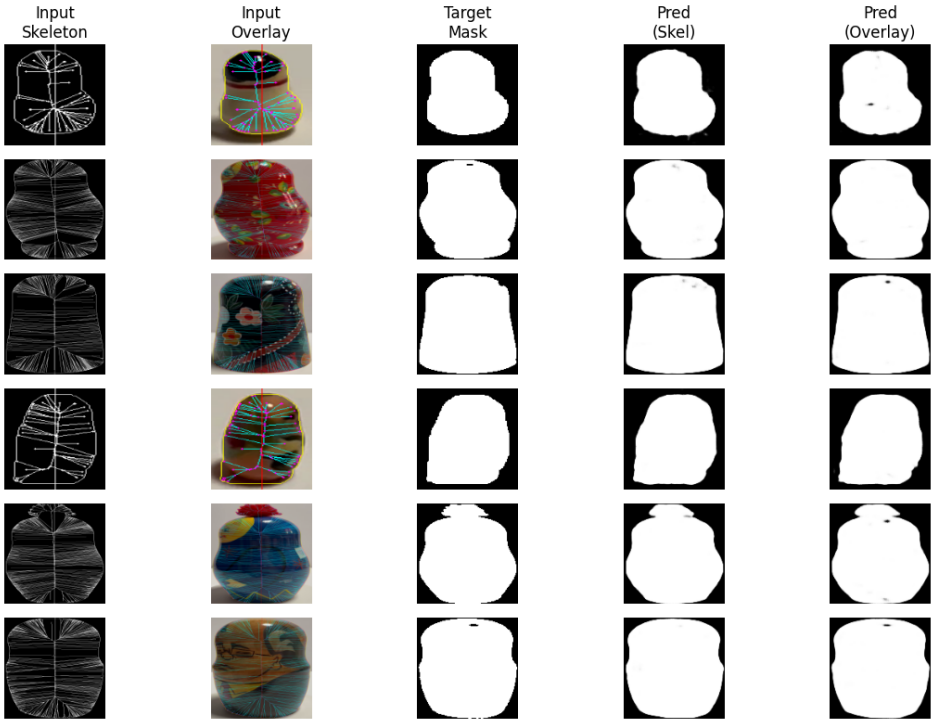


Figure 25. Reconstruction examples. Columns: *Input Skeleton*, *Input Overlay*, *Target Mask*, *Pred (Skel)*, *Pred (Overlay)*. Both regimes recover global shape reliably, with skeleton-only performing especially strongly on clean topology.

 doi:10.20944/preprints202602.0036.v1

3.9. 2D-to-3D Reconstruction of Matryoshka Dolls

Our 3D reconstruction pipeline converts short turntable videos of a Matryoshka doll into normalized 3D shape representations suitable for geometric analysis and learning. The process consists of four main steps: (1) video frame extraction, (2) 2D silhouette estimation and skeletonization via the Blum Medial Axis (BMA), (3) axisymmetric surface-of-revolution to produce a triangulated 3D mesh and sampled point cloud, and (4) optional surface refinement (e.g., Poisson reconstruction) when a watertight mesh is required. Finally, meshes belonging to a Matryoshka family are normalized and scaled into a shared coordinate system to form an explicit nested 3D representation.

Figure 26 summarizes all stages for a political Matryoshka family (IDs 4799–4805). The first row shows example frames, the second row the corresponding 3D point cloud projections, the third row the mesh projections, and the last row the 3D views of the reconstructed meshes.

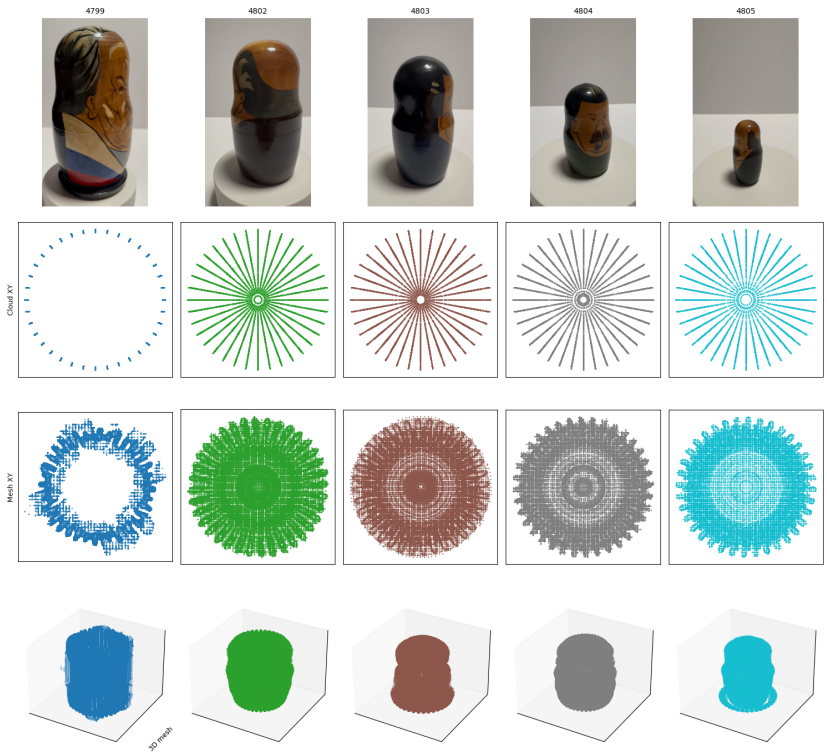


Figure 26. Grid visualization of the Matryoshka 3D reconstruction pipeline (top to bottom): original frames, point–cloud XY projections, mesh XY projections, and 3D mesh views.

For each video v, we extract frames at a fixed rate of `FPS = 2` until a maximum of `MAX_FRAMES = 20` frames is reached. Each frame is resized to a fixed processing width W_{proc} while preserving aspect ratio:

$$\text{scale} = \frac{W_{\text{proc}}}{W_{\text{orig}}}, \qquad (x, y)_{\text{new}} = \text{scale}\,(x, y)_{\text{orig}}. \tag{4}$$

The resized grayscale image $I \in [0, 255]^{H \times W}$ is binarized by Otsu thresholding. If the foreground and background are inverted, the mask is flipped such that the doll is white on black.

From the binary mask we extract the outer contour C using *cv2.findContours*. Let the contour points be

$$C = \{(x_k, y_k)\}_{k=1}^{N}. \tag{5}$$

We embed the contour into the complex plane,

$$z_k = x_k + i y_k, \qquad k = 1, \ldots, N, \tag{6}$$

 doi:10.20944/preprints202602.0036.v1

and compute a Delaunay triangulation over the point set $\{z_k\}_{k=1}^{N}$. For each oriented triangle $(u, v, w) \in \mathbb{C}^3$ the BMA is approximated by computing the circumcenter m and radius r:

$$d = (u - w)\,\overline{(v - w)}, \tag{7}$$

$$m = \frac{1}{2}\Big(u + v + i\,(u - v)\frac{\Re(d)}{\Im(d)}\Big), \tag{8}$$

$$r = |u - m|. \tag{9}$$

Only triangles whose orientation lies inside the object are kept,

$$\Im(d) > 0, \tag{10}$$

yielding a set of medial points $\{m_j\}$ and associated radii $\{r_j\}$. Very small radii are discarded as noise using a relative threshold

$$r_j > \alpha \max_k r_k, \tag{11}$$

with $\alpha = 0.05$ in our experiments.

The resulting skeleton is overlaid on the binary mask and written to disk. An example is shown in Figure 27.

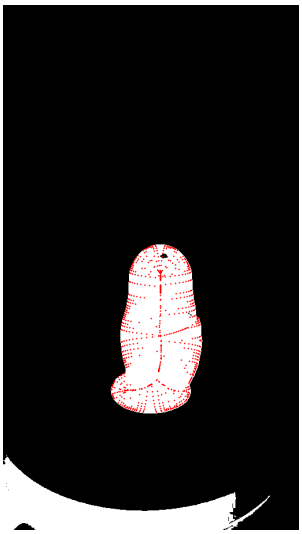


Figure 27. BMA skeleton extracted from a thresholded doll silhouette. The skeleton follows the internal symmetry of the doll and defines the centers and local radius of the generative surface.

To produce a 3D shape from a 2D silhouette, we treat the silhouette as an axisymmetric object and revolve its radius profile around an estimated vertical symmetry axis. Let H and W denote image height and width, and let x_0 be the estimated symmetry axis in the cropped ROI (obtained from the pipeline as `axis_x_crop`). For each image row y, we compute the silhouette radius

$$r(y) = \max_{x \in \mathcal{F}(y)} |x - x_0|, \tag{12}$$

where $\mathcal{F}(y)$ denotes the set of foreground pixels in row y. After optional smoothing of $r(y)$, we sample azimuth angles $\theta_k \in [0, 2\pi)$ for $k = 1, \ldots, N_\theta$. Each row generates a 3D ring:

$$\begin{aligned} X_{yk} &= r(y)\cos\theta_k, \\ Z_{yk} &= r(y)\sin\theta_k, \\ Y_{yk} &= \hat{y}, \end{aligned} \tag{13}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 February 2026 doi:10.20944/preprints202602.0036.v1

where $\hat{y}$ is a normalized vertical coordinate (optionally scaled to metric units). Connecting adjacent rings yields a triangulated mesh. We further sample points uniformly from the mesh surface and store them with face-normal estimates, producing point clouds appropriate for 3D learning (e.g., PointNet/DGCNN). This axisymmetric reconstruction is robust, lightweight, and does not require multi-view correspondence.

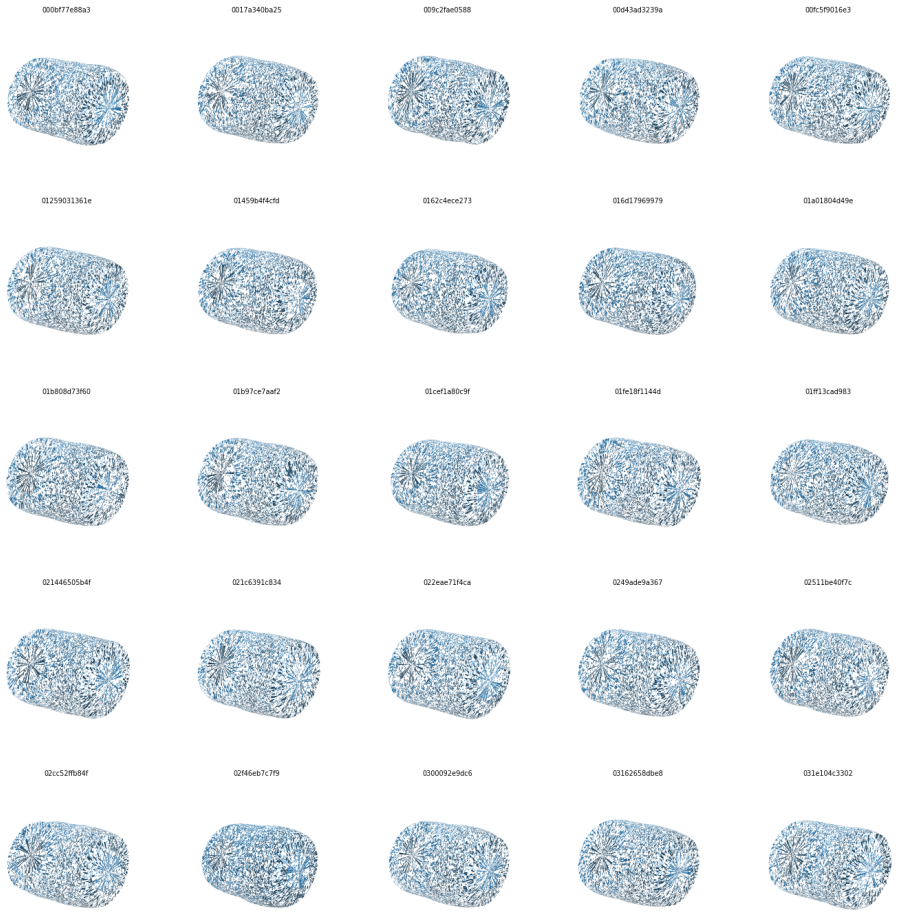


Figure 28. Axisymmetric 3D mesh generation. Each mesh is obtained by revolving the silhouette radius profile around the estimated symmetry axis, producing surfaces that can be directly sampled into point clouds for downstream 3D training.

Figure 29 shows a representative 3D point cloud and its characteristic "onion" structure induced by ring sampling. To obtain a watertight 3D mesh when required, we apply Poisson surface reconstruction to each point cloud. Let $\mathbf{p}_i \in \mathbb{R}^3$ denote the sample points and $\mathbf{n}_i$ their estimated normals. Poisson reconstruction solves for an indicator function χ such that

$$\nabla \cdot \nabla \chi = \nabla \cdot \mathbf{v}, \qquad \mathbf{v}(\mathbf{x}) = \sum_i \mathbf{n}_i \, \delta(\mathbf{x} - \mathbf{p}_i), \tag{14}$$

and extracts the zero level set of χ as the reconstructed surface. In practice, we use Open3D's implementation with octree depth `depth = 8` and a scale factor of `1.1`. Low-density vertices are removed by discarding points whose Poisson density falls below the 2 % quantile. The resulting triangle meshes are saved. For each mesh we also generate 2D orthographic projections (XY, XZ, YZ) and 3D views as shown in the bottom rows of Figure 26.

 doi:10.20944/preprints202602.0036.v1

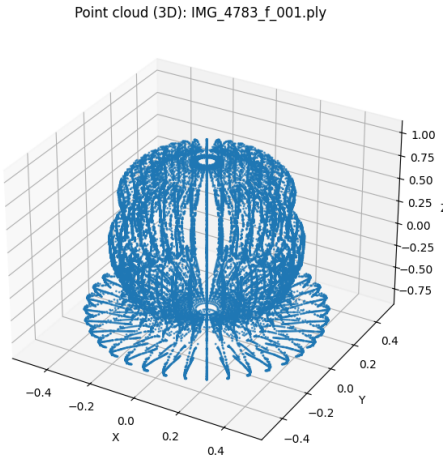


Figure 29. 3D point cloud generated by revolving a silhouette-derived profile around the vertical axis. Rings at different heights form an axisymmetric sampling of the outer shell of the doll.

To visualize a full Matryoshka family as a nested 3D object, we load all meshes whose filenames belong to a political Matryoshka family (target video IDs: *4799*, *4802*, *4803*, *4804*, *4805*). For each mesh, we compute centered vertices and a characteristic radius

$$\mathbf{v}_i' = \mathbf{v}_i - \bar{\mathbf{v}}, \qquad R = \max_i \|\mathbf{v}_i'\|_2, \tag{15}$$

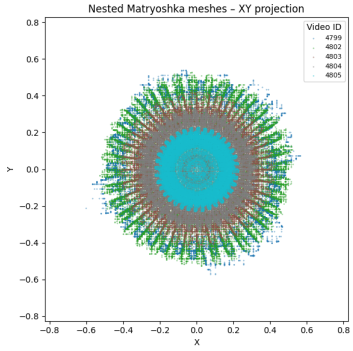
and sort dolls from outer to inner by decreasing R. We then assign target radii

$$R_{\text{target}}^{(k)} \in [0.4, 1.0], \qquad k = 1, \ldots, K, \tag{16}$$

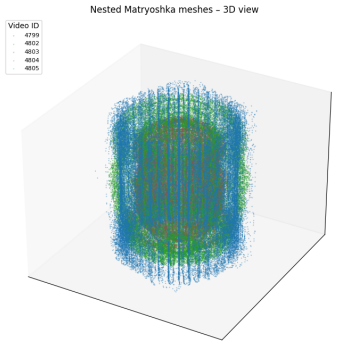
computed as a linear spacing between 1.0 (outer) and 0.4 (inner). Each mesh is uniformly scaled by

$$s_k = \frac{R_{\text{target}}^{(k)}}{R^{(k)}}, \qquad \tilde{\mathbf{v}}_i^{(k)} = s_k \, \mathbf{v}_i'^{(k)}, \tag{17}$$

which guarantees that all dolls are centered and strictly nested without intersections. The nested XY projection and nested 3D view are shown in Figures **??** and **??**, respectively. Color encodes the video ID and, therefore, the physical doll in the set. The cylindrical envelope corresponds to the shared turntable radius, while vertical variation reflects height and head/torso shape differences between dolls.



(**a**) XY projection. Colors denote video IDs, ordered from outer to inner.



(**b**) 3D view. The dolls share a common center and axis; only scale changes between shells.

Figure 30. Nested Matryoshka meshes shown in (**a**) XY projection and (**b**) full 3D view.

 doi:10.20944/preprints202602.0036.v1

Table 9 lists the key tunable parameters for frame extraction and pre–processing. Table 10 summarizes the parameters used in skeleton filtering, axisymmetric revolution, and Poisson reconstruction. Table 11 contains visualization parameters for the plots.

Table 9. Acquisition and pre–processing parameters.

Symbol	Name	Value	Description
FPS	Frame rate	2	frames extracted per second
M	*MAX_FRAMES*	20	max. frames per video
W_{proc}	*IMG_WIDTH*	512	processing width (pixels)
	Upscale factor	1	contour precision scaling

Table 10. Geometry, skeleton and reconstruction parameters.

Name (Symbol)	Value	Description
radius threshold (α)	0.05	discard medial points with $r_j \leq \alpha \max_k r_k$
num_angles (N_θ)	36	azimuth samples for surface of revolution
Poisson depth (d)	8	octree depth in Poisson reconstruction
Poisson scale (s)	1.1	reconstruction bounding–box scale factor
density quantile (q)	0.02	vertices below q-quantile are removed
min. nested radius (R_{min})	0.4	target radius of smallest doll
max. nested radius (R_{max})	1.0	target radius of outermost doll

Table 11. Visualization parameters.

Name (Symbol)	Value	Description
max_points (N_{pc})	40000	subsample of point–cloud points
marker size (2D)	0.2–0.3	scatter size for XY projections
marker size (3D)	0.2–0.3	scatter size for 3D views
colormap	tab10	distinct color per doll/video ID

3.10. 3D Multitask Mesh Pipeline for Matryoshka Authentication and Semantic Classification

This section presents an integrated evaluation of the **3D branch** of the Matryoshka benchmark, which operates on reconstructed meshes and sampled point clouds to quantify how much semantic and authenticity signal is recoverable from *geometry alone* and how multimodal grounding affects representation quality. We report: (i) a controlled **reconstruction benchmark** comparing a direct GT-mask pipeline (A) against a learned skeleton-to-silhouette surrogate (B), and (ii) a **150-epoch multimodal multitask study** over modern point-cloud backbones under Late Fusion with Qwen textual embeddings, complemented by latent-space visualizations.

Each reconstructed mesh is converted into a fixed-size point set to standardize downstream learning across backbones. Concretely, for every mesh we sample a surface point cloud of size $N = 2048$ (surface sampling when possible; otherwise falling back to vertices with padding). Each point cloud is then centered and scaled to the unit sphere, producing an input tensor of shape $(3, N)$ suitable for point-based backbones. This normalization removes trivial scale/translation cues and makes comparisons across reconstruction variants and networks more interpretable. Supervision is transferred consistently from the 2D metadata: each mesh inherits both (i) a fine-grained semantic label (Style) and (ii) a coarse authenticity label (RU / non-RU / unknown) associated with the source capture. This creates a controlled multitask protocol in which style tests fine-grained discrimination from geometry, while authenticity tests whether provenance-related signatures remain observable after reconstruction and point sampling.

We first test whether a learned silhouette surrogate can replace ground-truth masks in large-scale generation without degrading geometric fidelity. The benchmarking suite compares:

 doi:10.20944/preprints202602.0036.v1

- **Pipeline A (GT Mask):** direct reconstruction from ground-truth silhouettes.
- **Pipeline B (Skel-AE):** skeleton → convolutional autoencoder (AE) → predicted silhouette → reconstruction.

Table 12 consolidates space–time efficiency and geometric fidelity at $N = 2000$. Pipeline B achieves **near-lossless reconstruction parity**, with **Avg IoU** $= 0.979$ and **Avg Dice** $= 0.989$, while also being **approximately 5% faster** in generation time (8.586s vs. 8.995s). These results justify using Pipeline B as a practical surrogate in extended training runs and ablations, especially when GT masks are unavailable, expensive to curate, or intentionally withheld to validate end-to-end robustness.

Table 12. Consolidated Space-Time Efficiency and Geometric Fidelity Summary ($N = 2000$).

Pipeline	Gen Time (s)	Mesh Size (MB)	Avg IoU	Avg Dice
A (GT Mask)	8.995	0.606	–	–
B (Skel-AE)	8.586	0.607	0.979	0.989

After establishing that Pipeline B preserves geometry with near-lossless fidelity, we report a large-scale multimodal learning experiment conducted over **150 epochs**. We benchmark five advanced 3D backbones under a **Late Fusion** regime that integrates point-cloud geometry with **Qwen-2.5-0.5B** textual embeddings. The late-fusion design isolates the representational contribution of geometry (via the 3D encoder) while allowing text to stabilize and disambiguate semantic boundaries through a fusion head. The extended 150-epoch horizon is intentionally challenging: training dynamics frequently plateau or destabilize without adaptive learning-rate control. A `ReduceLROnPlateau` scheduler is therefore used as a stability mechanism, enabling continued progress past plateaus and reducing the risk of late-epoch divergence.

Table 13 summarizes performance on (i) **Style** (160 classes) and (ii) **Authenticity** (3 classes). Two consistent findings emerge:

1. **Authenticity is systematically more learnable than fine-grained style from geometry**, even under multimodal grounding. This is consistent with authenticity correlating with coarser geometric cues (global proportions, silhouette curvature, base/top profiles) that survive reconstruction noise.
2. **Modern architectures exhibit improved stability and sensitivity.** In particular, **PointMLP** achieves the strongest authenticity accuracy (**0.4825**), while **PCT** yields the best style accuracy (**0.0475**), indicating that transformer-style global modeling is comparatively beneficial for fine-grained semantic structure, whereas MLP-style point mixing provides strong robustness for the provenance-oriented task.

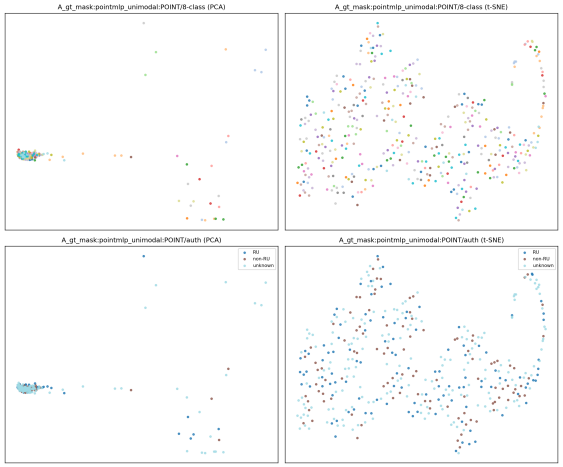
Table 13. Backbone Performance Metrics for Style (160 Classes) and Authenticity (3 Classes) Tasks (150 epochs).

Backbone	Fusion	Params	Auth Acc	Style Acc
PointNet	Late	0.38M	0.4675	0.0450
DGCNN	Late	0.56M	0.4650	0.0325
PointNet++	Late	1.80M	0.4775	0.0425
PCT	Late	1.07M	0.4800	**0.0475**
PointMLP	Late	1.07M	**0.4825**	0.0425

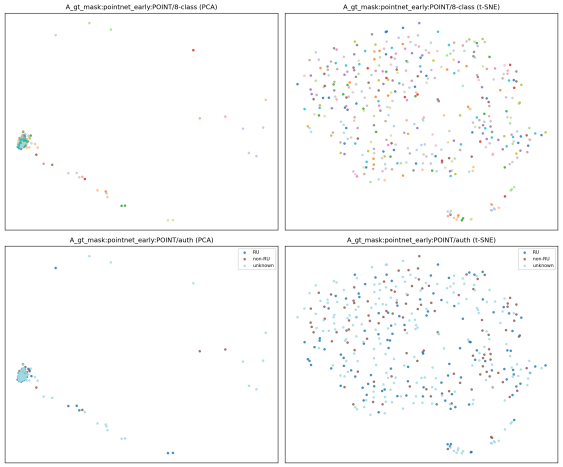
The authenticity task has a chance baseline of approximately $1/3 \approx 0.333$, while the 160-way style task has a chance baseline of $1/160 = 0.00625$. Thus, authenticity results reflect substantial recoverable signal, whereas style remains difficult in the geometry-centric regime even with multimodal fusion, indicating persistent overlap among fine-grained categories when constrained to reconstructed shape cues. To interpret why authenticity is more learnable than fine-grained style, we visualize the learned

 doi:10.20944/preprints202602.0036.v1

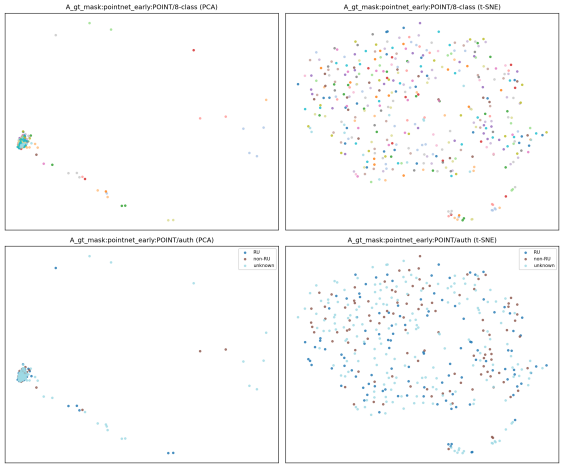
embeddings on the validation split using PCA and t-SNE projections. Figure 31 includes your consolidated 2×2 panels (top row: style coloring; bottom row: authenticity coloring) and is used as qualitative evidence of representational separability. Across backbones, the authenticity coloring exhibits more coherent regional structure than the 160-way style coloring, reinforcing that fine-grained semantic boundaries are heavily entangled in geometry space, while provenance signals remain more stable. This qualitative latent structure is consistent with Table 13, where style accuracy stays low relative to authenticity.



(**a**) PointMLP (Unimodal) latent map



(**b**) PointNet (Early Fusion) latent map



(**c**) PointNet (Early Fusion) latent map (alt)

Figure 31. Latent-space projections (PCA / t-SNE) colored by style (top) and authenticity (bottom), using exported 2×2 panels with exact filenames preserved.

We retain the 2D-vs-3D family comparison to contextualize geometry-centric learning against visual baselines (Figure 32). The integrated conclusion remains consistent across the full benchmark: **authenticity is more recoverable from 3D geometry than fine-grained semantic style**, motivating multimodal fusion as a stabilizing mechanism and motivating higher-fidelity geometry or additional cues (texture, paint patterns, inscriptions) as the primary path to improving 160-way style recognition.

 doi:10.20944/preprints202602.0036.v1

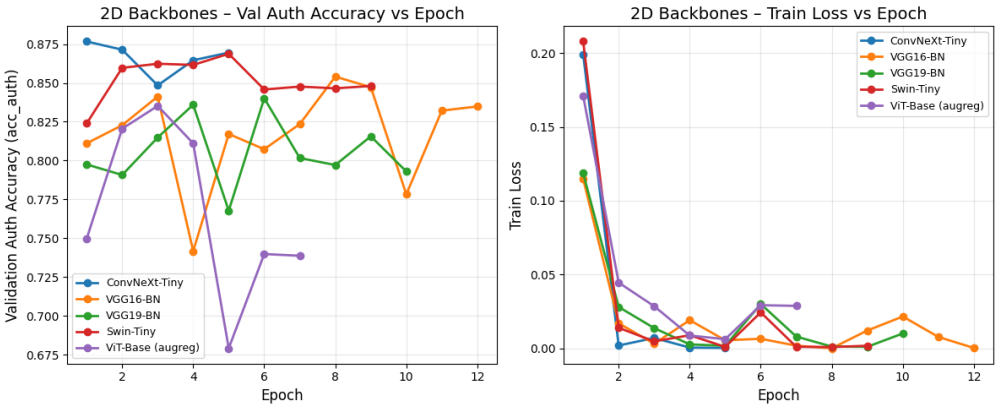


Figure 32. Comparison of 2D and 3D family dynamics on the Matryoshka multitask protocol.

Lessons learned: The reconstruction benchmark validates **Pipeline B (Skel-AE)** as an efficient and near-lossless surrogate for GT-mask reconstruction (Dice = 0.989). The 150-epoch multimodal study shows that **modern 3D architectures (PointMLP, PCT)** yield the most stable long-run behavior and the strongest endpoint metrics, but that **fine-grained style remains intrinsically challenging** from geometry-centric representations. Latent-space projections corroborate these trends: authenticity exhibits more coherent structure than style, aligning with the measured gap between 3-way and 160-way performance.

3.11. Zero Shot Point Cloud Completion

The zero-shot point cloud completion pipeline that we are integrating here is a critical component because nested Matryoshka dolls and turntable capture often yield *partial* 3D geometry (self-occlusions, missing back-surfaces, interior cavities, and sparse coverage). Rather than treating incompleteness as a failure case, we explicitly integrate a completion stage that can *synthesize missing geometry* without requiring paired "partial-to-complete" Matryoshka training data. The integration point within the full system is shown in Figure 35. The way this part of the pipeline works is as follows. The method consists of three main stages: (i) depth-map synthesis, (ii) generative mesh reconstruction, and (iii) rigid registration. The depth-map synthesis stage is responsible for generating a plausible depth map from the partial point cloud. This step is crucial as it serves as the control signal to the image generative model to produce an RGB image. However, camera intrinsics and extrinsics are not available in the partial point cloud, thus we need to first estimate a camera pose that captures the most amount of depth information. We achieve this via a two-stage pose estimation process: convex-hull visibility initialization and depth image refinement. Optionally, when the depth image is too sparse or not ideal, we can apply a depth inpainting step to fill the holes in the depth map. The depth-map synthesis stage is responsible for generating a plausible depth map from the partial point cloud. This step is crucial as it serves as the control signal to the image generative model to produce an RGB image. However, camera intrinsics and extrinsics are not available in the partial point cloud, thus we need to first estimate a camera pose that captures the most amount of depth information. We achieve this via a two-stage pose estimation process: convex-hull visibility initialization and depth image refinement. Optionally, when the depth image is too sparse or not ideal, we could also apply a depth inpainting step to fill the holes in the depth map.

We first uniformly sample a set of camera origins $\mathbf{o}_C$ on a sphere of radius R centered at the centroid $\mathbf{o}_P$ of the partial point cloud $\mathbf{P}$. The camera angles are defined by the vector $\mathbf{o}_C - \mathbf{o}_P$. Then for each $\mathbf{o}_C$, we use a sphere of radius R centered at $\mathbf{o}_C$ to compute the spherical mirrored point cloud $\hat{\mathbf{P}}$ with

$$\hat{p}_i = F(p_i) = p_i + 2(R - \|p_i\|)\frac{p_i}{\|p_i\|}$$

for each point p_i in the partial point cloud $\mathbf{P}$. The convex hull of the concatenated set $\{\hat{P}, \mathbf{o}_C\}$ is computed, and the camera with the most vertices on the convex hull is selected as the initial camera

 doi:10.20944/preprints202602.0036.v1

pose $\mathbf{o}_I$. According to [51], this step is crucial as it ensures that the depth map is on the right hemisphere. We visualized the normalized depth pixel count and the convex hull vertices count as the visibility in Figure 33. However, we found that this step is not sufficient to ensure a good depth map, as several camera poses around the optimal point can lead to convex hulls with the same number of vertices. We then refine the depth camera pose by rendering the depth map from the initial camera pose and computing the visibility of pixels. The visibility is computed as the total number of pixels that are visible from the camera. Starting from the initial camera pose $\mathbf{o}_I$, we uniformly sample a set of camera poses $\mathbf{o}_C$ on a tangential disk with shrinking radius $R_i = 2^{-i/4}R$ centered at $\mathbf{o}_I$. For each camera pose, we render the depth map and compute the visibility. The camera pose with the most visible pixels is selected as the final camera pose $\mathbf{o}_F$, and we iterate this process until convergence. The depth map generated from the above two steps may still be sparse or contain holes. To address this issue, we generate a low-resolution depth map and a high-resolution depth map from $\mathbf{o}_F$. Then we compute the binary XOR mask of the two depth maps and use the mask to inpaint the low-resolution depth map with a diffusion model. We use the Stable Diffusion 2 Inpainting model with 100 inference steps and a guidance scale of 2.0.

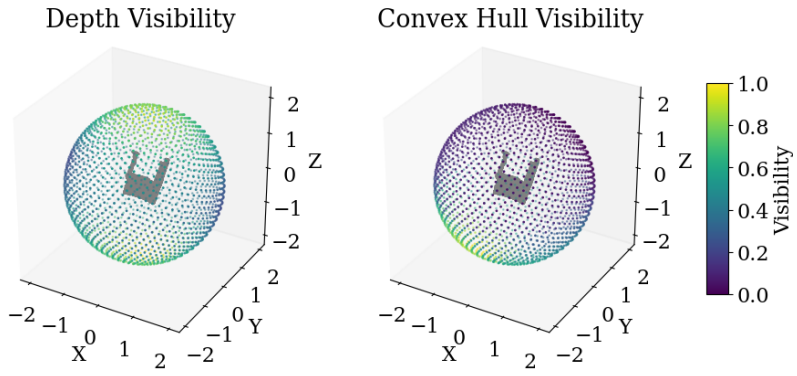


Figure 33. The normalized convex hull vertices count and the normalized depth pixel count as the visibility. Two modes can be observed on the sphere using the depth visibility method, which may cause depth inversion by choosing the wrong hemisphere. The convex hull visibility only has one mode, which is the front side of the partial point cloud, thus avoids the depth inversion problem.

We use the depth map generated from the previous stage into a controlnet of the diffusion model to generate a RGB image. To be specific, we use the Stable Diffusion v1.5 model with a controlnet that is trained on the depth map. The depth conditioning scale is set to 1.1 to ensure geometric consistency. The generated RGB image is then used as the input to the 3D generative model. We use the Hunyuan3D model to generate a mesh from the RGB image. The model is a latent diffusion model with an implicit surface VAE trained on a large-scale dataset of 3D meshes. The model is capable of generating high-quality meshes from single-view images. We use the default settings of the model with a guidance scale of 7.5 and 50 inference steps. The generated mesh is then converted to a point cloud by sampling points on the surface of the mesh. We use the Poisson disk sampling method to sample points on the surface of the mesh. The sampled points $\mathbf{P}_G$ are then used as the initial point cloud for the next stage. The generated point cloud from the previous stage may not be aligned with the original partial point cloud. To address this issue, we use a rigid registration method to align the two point clouds. In [51], the authors adopt the Point-to-Plane ICP algorithm [64] to align the two point clouds. The algorithm iteratively refines the rotation and translation of the generated point cloud to minimize the distance to the original partial point cloud. The algorithm converges when the distance between the two point clouds is below a certain threshold. However, we found that the Point-to-Plane ICP algorithm is sensitive to the initial alignment of the two point clouds, and thus does not lead to stable convergence. Therefore, we adopt a novel registration method that is based on Gradient Descent to search for the optimal rotation, translation, and scaling of the generated point cloud.

 doi:10.20944/preprints202602.0036.v1

We aim to optimize the Euler rotation angles α, β, γ, the translation vector $\mathbf{t}$, and the scaling factor s of the generated point cloud $\mathbf{P}_G$ to minimize the Unidirectional Chamfer Distance (UCD) from the original partial point cloud $\mathbf{P}_P$ to the generated point cloud $\mathbf{P}_G$. The UCD is defined as:

$$\text{UCD}(\mathbf{P}_P, \mathbf{P}_G) = \frac{1}{|\mathbf{P}_P|} \sum_{p_i \in \mathbf{P}_P} \min_{g_j \in \mathbf{P}_G} \|p_i - g_j\|_2$$

where $|\mathbf{P}_P|$ is the number of points in the partial point cloud. In order to make the process differentiable, we construct the transformation matrix T as:

$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & s \end{bmatrix}$$

where $\mathbf{R}$ is the rotation matrix constructed from the Euler angles α, β, γ. The transformed point cloud $\mathbf{P}'_G$ is then transformed as:

$$\mathbf{P}'_G = \mathbf{T}\mathbf{P}_G.$$

The optimization problem can then be formulated as:

$$\min_{R,t,s} \text{UCD}(\mathbf{P}_P, \mathbf{P}'_G).$$

Overall, this part of the pipeline is not supposed to work on all other kinds of data. There are some limitations in this which are due to the stochastic nature of the generative pipeline, the generated point cloud may not be exactly aligned with the original partial point cloud. Moreover, since our registration method is based on gradient descent, which is more efficient but may get stuck in local minima due to the nature of the optimization algorithm, there is a risk that the final generated shape is not perfectly aligned with the partial point cloud. For example, here, we evaluate a geometry-only zero-shot point cloud completion baseline that operates without category supervision, training data, or learned shape priors. Given a partial point cloud acquired from a single viewpoint, the method estimates a dominant viewing direction and projects the observed geometry into a depth representation. A completed spatial envelope is then obtained by back-projecting the visible depth values into 3D space, yielding an expanded point set that is geometrically consistent with the observed surface. As shown in the figure here.

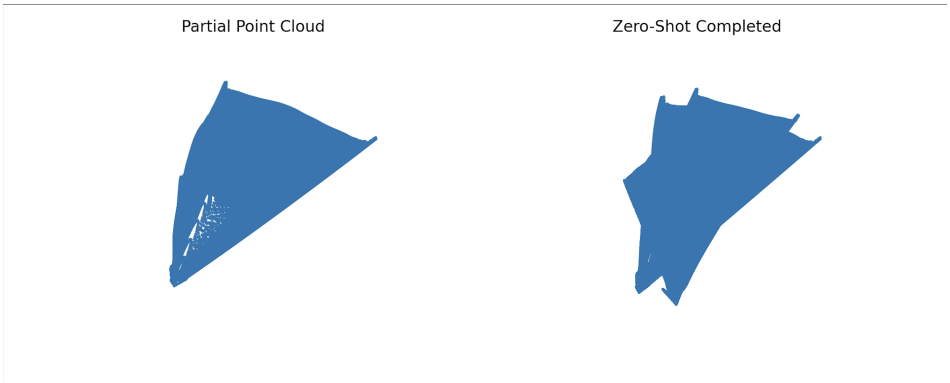


Figure 34. Geometric zero-shot point cloud completion on a Matryoshka doll frame

The completion extends the partial observation by inferring plausible occupied regions behind the visible geometry. Since the method does not incorporate semantic knowledge or object-level priors, the reconstructed shape reflects surface continuity rather than object-specific structure (For example, symmetry or category-dependent geometry). This baseline highlights both the feasibility and inherent limitations of purely geometric zero-shot completion and serves as a reference point for comparison with learned and diffusion-based 3D completion methods. For example, like here in this case: the zero-

 doi:10.20944/preprints202602.0036.v1

shot geometric completion method operates without category supervision, training data, or learned shape priors. The algorithm projects the observed 3D points into a depth representation along the dominant viewing direction and subsequently back-projects this depth to estimate occluded regions, yielding a completed spatial envelope. As shown in the figure, the method successfully extrapolates a dense, contiguous geometry that is strictly consistent with the visible surface geometry in the input. Importantly, the reconstruction prioritizes geometric continuity and visibility consistency rather than semantic plausibility. As a result, the completed shape reflects the maximal occupied volume compatible with the single-view observation, leading to elongation effects aligned with the viewing frustum. Object-specific structural cues - such as the known axial symmetry and smooth curvature of a Matryoshka doll - are not explicitly recuperated, as the method does not incorporate semantic or learned priors. Nevertheless, this experiment demonstrates that purely geometric zero-shot completion can recover meaningful and stable 3D structure from extremely limited input, providing a strong data-agnostic baseline. These results highlight both the strengths and limitations of geometry-only inference and motivate the integration of learned 3D generative priors (For example diffusion-based models) like Hunyuan3D to further refine semantic plausibility and global shape consistency which will be used in this pipeline in the upcoming future to further improve upon the zero shot point cloud completion using the prior 3D Diffusion models on our novel Matryoshka Dolls Dataset.

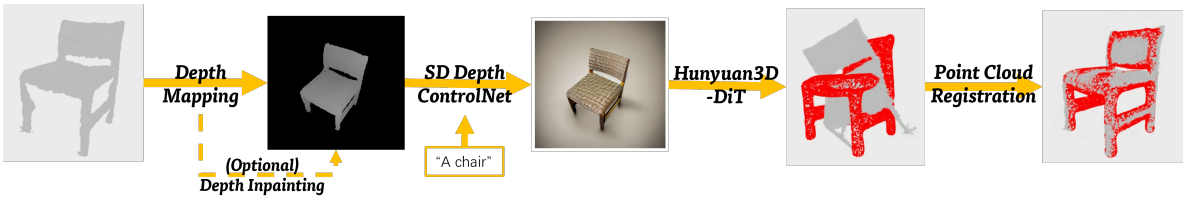


Figure 35. End-to-end Matryoshka pipeline integrating 2D recognition, text modality, skeleton/BMA extraction, and 3D reconstruction.

Our completion pipeline is designed to operate in a *calibrated* multimodal setting. Given a set of high-quality Matryoshka images, we run COLMAP to reconstruct a sparse (or dense) point cloud and recover accurate camera poses and intrinsics. Using these calibrated poses, we project 2D semantic features (obtained from the trained 2D/multimodal encoders) onto the reconstructed 3D points. This produces an aligned 2D–3D feature space that supports downstream 3D reasoning under text prompts. In particular, the aligned representation enables a generative 3D model to condition on both the partial geometry and the text instruction, and synthesize the missing geometry required to complete the point cloud. I was also able to obtain some results by processing the codebase and

A key requirement for reliable feature projection (and therefore reliable completion) is strong photogrammetric calibration. In our current experiments, COLMAP achieves sub-pixel mean reprojection error, which is typically considered excellent reconstruction quality. We also track the total number of registered images and optimization statistics (bundle adjustment residuals and total parameters) to ensure the geometry is sufficiently constrained before invoking the generative completion step.

While the integration is functional, several challenges remain central to making zero-shot completion robust in this cultural-heritage setting: (i) feature aggregation robustness when projecting from high-resolution 2D imagery onto point clouds of varying density, (ii) domain gap between 2D models trained on generic web imagery and the photogrammetric capture distribution, (iii) physical plausibility and seamless stitching of newly generated geometry with the existing sparse reconstruction, and (iv) the significant time and memory cost of the end-to-end pipeline (SfM, feature extraction passes, and generative 3D synthesis).

Some of the results can be found in Figure 34.

 doi:10.20944/preprints202602.0036.v1

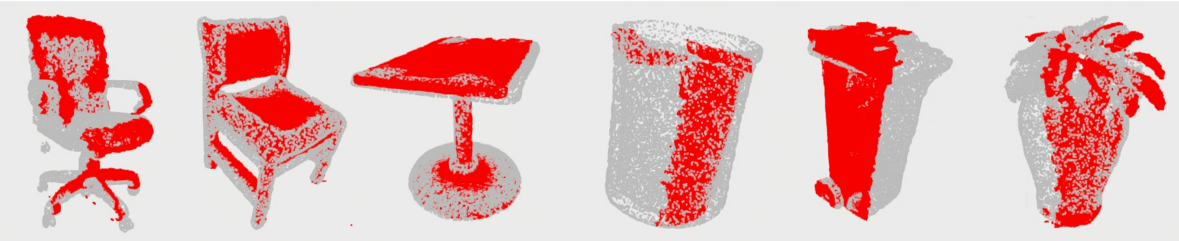


Figure 36. Qualitative results on the Redwood dataset. The partial point cloud is in red, the inferred point cloud is in gray. Our method is capable of inferring reasonable missing parts of the object and fits the original shape.

Table 14. Quantitative results on the Redwood [59] dataset. We report the l_2 Chamfer Distance (CD) in units of 10^{-2} (lower is better). Best results are in **bold**.

Method	Speed	Zero-shot	Table	Swivel Chair	Arm-Chair	Chair	Sofa	Vase	Off-Can	Vespa	Wheelie-Bin	Tricycle
PointTr [50]	–	No	1.86	4.08	1.95	2.69	2.96	4.05	4.82	2.00	2.78	**1.70**
SDS-Complete [59]	∼25 h	Yes	1.35	1.96	2.18	2.77	2.95	3.00	3.79	3.36	2.69	3.18
ComPC [54]	∼15 min	Yes	1.67	**1.04**	**1.28**	1.42	–	2.94	3.51	**1.39**	–	2.42
Ours	**∼1 min**	Yes	**1.05**	1.53	1.40	**1.17**	**2.65**	**2.38**	**3.07**	2.76	**2.61**	2.64

4. Practical Applications

The multitask 2D–3D models jointly predict an 8-class semantic label and a 3-class authenticity label. In practice, these models can serve as a triage tool for collectors, auction houses, and museums: new dolls are photographed or scanned, passed through the pipeline, and suspicious or ambiguous items are flagged for expert review (Figure 37). Beyond market authenticity, the system helps safeguard traditional values embedded in genuine Russian schools (e.g., Sergiev Posad, Semenov, Polkhov-Maidan) by distinguishing them from mass-produced or culturally diluted replicas. This is particularly important for preserving religious iconography and folk motifs that carry strong cultural and spiritual meaning.

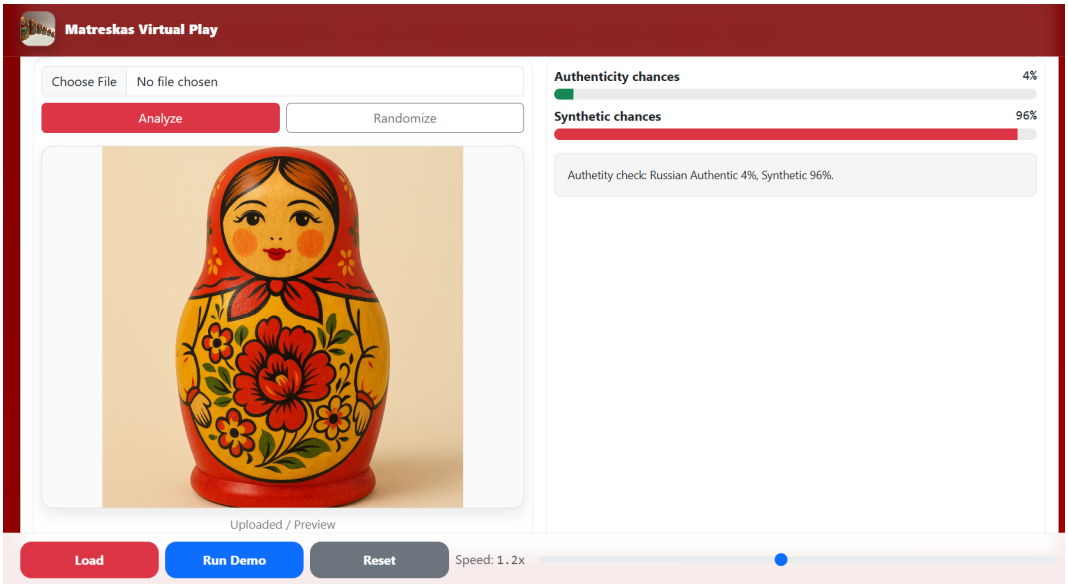


Figure 37. Prototype illustrating the operational triage output of the proposed pipeline, including calibrated probabilities for authenticity vs. synthetic likelihood.

The dedicated religious and political classes allow researchers to systematically study how Matryoshkas encode religious narratives, symbolism, and cultural identity. By aggregating predictions

 doi:10.20944/preprints202602.0036.v1

and captions across large collections, the pipeline can reveal patterns in how religious figures (saints, icons, biblical scenes) and cultural tropes are represented across time, regions, or political periods. This provides a quantitative complement to qualitative art-historical analysis and supports discussions around cultural appropriation, respectful representation of faith, and the evolution of traditional values in modern merchandise and souvenirs.

Marketplaces hosting "Russian nesting doll" listings are vulnerable to misleading descriptions and low-quality replicas. Our 2D frame-level classifiers and 3D mesh-based backbones can be integrated as a pre-screening step at upload time: items classified as non_authentic, merchandise, or non_Matryoshkas can be down-ranked, flagged, or forced to carry "replica/novelty" labels. For explicitly religious or political dolls, platforms may choose to apply different moderation policies or region-specific regulations. This improves transparency for buyers and protects both traditional craft communities and religious communities from misrepresentation.

Because the models operate at the level of videos, frames, 3D meshes, and explicit nested structure, they can support very rich digital catalog entries. For each set, the pipeline can (i) identify all nested dolls, (ii) map each 3D mesh to its semantic class and authenticity label, and (iii) reconstruct the full "nesting tree" in 3D. Curators can use these meshes to virtually open and close dolls, inspect inner layers that are physically fragile, and compare geometric and iconographic details across sets. For religious dolls, this enables non-invasive exploration of inner icons that might rarely be handled in physical collections. The same 3D infrastructure supports shape-based similarity search, geometry-aware clustering of schools, and longitudinal monitoring of physical degradation.

To translate these capabilities into an accessible user experience, we implemented a mobile-oriented web prototype consisting of two mini-games that visualize intermediate pipeline artifacts and the final nested outcome. Figure 38 shows **Game 1 (Single Doll Pipeline)**, which demonstrates the transformation from an input image with a coarse mask to an intermediate silhouette representation and then to a final output view. This interaction makes the segmentation and shape abstraction stages explicit, enabling users to understand what information the model consumes and produces at each stage. The prototype also supports a "Load / Run Demo / Reset" workflow and a speed control slider, allowing repeatable demonstrations in classroom and museum settings.

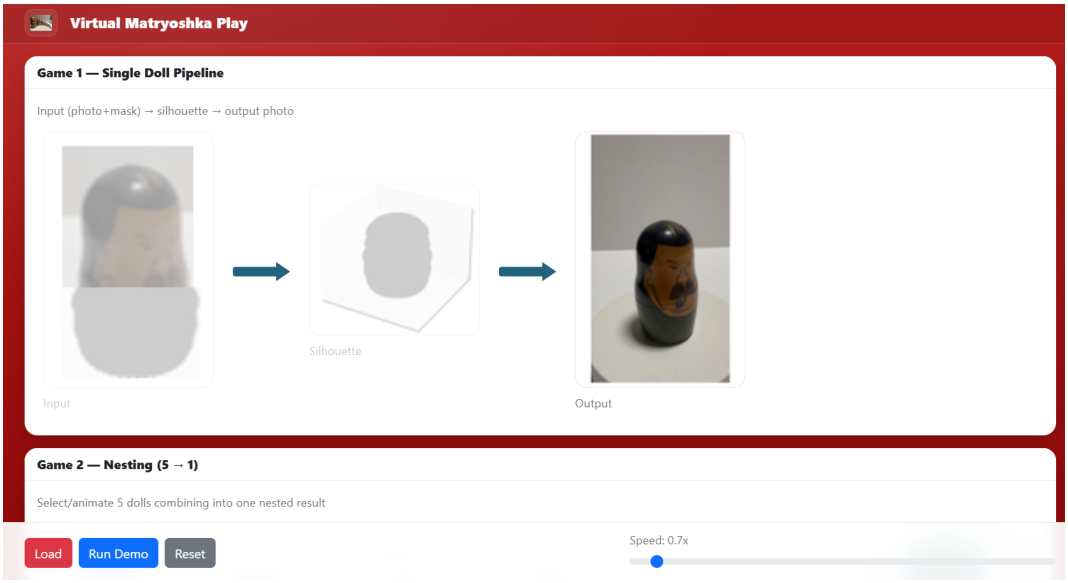


Figure 38. Web prototype result for Game 1 (Single Doll Pipeline): input (photo+mask) → silhouette → output, presented as a mobile-friendly, stepwise visualization.

Figure 39 presents **Game 2 (Nesting 5 → 1)**, where multiple doll representations (e.g., class- or layer-specific silhouettes/meshes) are visually combined into a single nested result. This design operationalizes the core concept of Matryoshka hierarchy: users can observe the left-to-right assembly

 doi:10.20944/preprints202602.0036.v1

process and the final consolidated structure. While the current version uses pre-rendered assets for clarity and stability on mobile devices, the interface is directly compatible with future integration of real-time 3D viewers and model-in-the-loop inference (e.g., generating silhouettes from uploaded photos, then producing predicted nesting order and overlays).

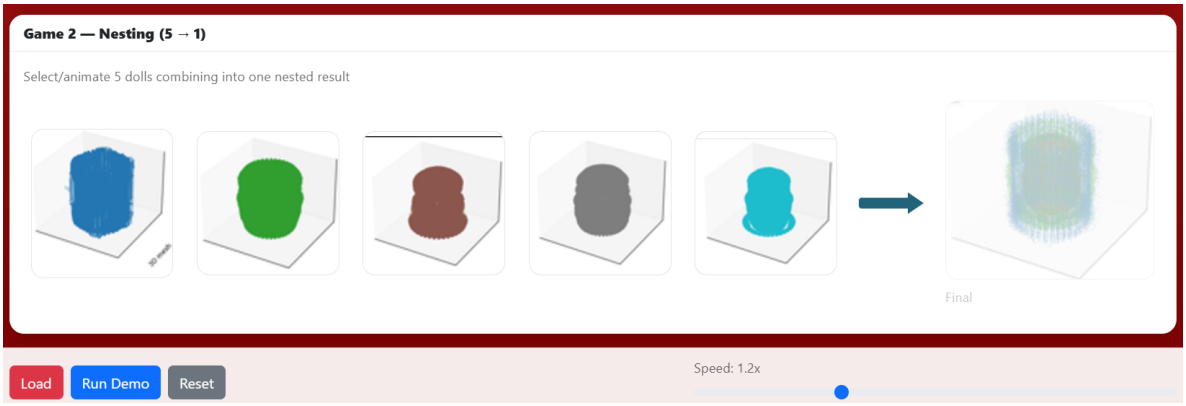


Figure 39. Web prototype result for Game 2 (Nesting 5 → 1): multiple dolls are assembled into a single nested structure through an animated left-to-right composition.

The reconstructed 3D meshes and nesting hierarchy make the dataset directly usable in virtual reality (VR) and augmented reality (AR). In a museum or classroom, visitors can "pick up" a virtual Matryoshka doll, rotate it, and virtually open it layer by layer, with overlays showing class predictions, authenticity scores, and textual descriptions. Religious and culturally significant scenes can be contextualized with narration or multilingual captions. In AR, users may point a tablet or headset at a physical doll and see an overlaid 3D avatar with school attribution, authenticity likelihood, and links to similar pieces in the collection. This turns a static object into an interactive, educational artifact while preserving the original from excessive handling.

The pipeline also supports playful, age-appropriate educational tools. Children can interact with virtual Matryoshkas in a game-like environment:

- **Classification games:** "Guess the school/class" or "authentic vs. replica," where the model provides feedback and short explanations.
- **Nesting puzzles:** drag-and-drop 3D meshes (or silhouettes) to reassemble a scrambled set into the correct nesting order, reinforcing spatial reasoning and cultural awareness.
- **Story mode:** for religious or culturally themed dolls, the system can present short stories or legends tied to the imagery on each layer.

Gamified scoring, badges, and "museum quests" can encourage children to explore art, religion, cultural identity, and traditional values in a respectful, engaging way while leveraging the same backbone models used for research. The current web prototype concretely demonstrates feasibility of this direction by showing that the nesting concept and key intermediate artifacts can be communicated effectively in a lightweight interface suitable for mobile devices.

From a machine learning perspective, the Matryoshka dataset and pipeline serve as a compact but realistic benchmark for multimodal, multitask learning under strong intra-class variation and label imbalance. Models must simultaneously reason about 2D appearance (painted motifs, colors, religious icons), 3D geometry (shape, nesting structure), and semantic labels (8-class plus authenticity). This makes the framework suitable for testing advanced 3D backbones (e.g., PointNet variants, 3D Swin, point transformers), fusion strategies (early/late/mid fusion of 2D, 3D, and text), and curriculum- or game-inspired training setups. Researchers can plug in alternative architectures and directly compare performance on standardized splits and metrics.

Wrapping the system into web, VR/AR, or mobile interfaces enables broader public engagement. Users worldwide can upload their own dolls, obtain probabilistic class and authenticity predictions, and read auto-generated descriptions that explicitly note uncertainty and limitations. This can spark

 doi:10.20944/preprints202602.0036.v1

cross-cultural conversations about what counts as "authentic," respectful religious depiction, and the role of traditional crafts in a globalized, AI-mediated world. In this way, the Matryoshka pipeline becomes not just a technical artifact, but a platform for exploring religious and cultural identity, traditional values, and multimodal AI through the lens of nested 2D–3D representations.

5. Results

This section reports the quantitative and qualitative outcomes of the proposed Matryoshka authentication benchmark across unimodal and multimodal evidence streams: (i) 2D appearance baselines, (ii) text-only caption baselines, (iii) 2D+text multimodal fusion, (iv) skeleton-based geometric compression, and (v) reconstructed 3D geometry learning and completion.

Table 4 summarizes the multi-task performance of five 2D backbones. Swin-Tiny achieves the strongest overall generalization, yielding the best test accuracy and macro-AUPRC across both the 8-way style task and the 3-way authenticity task, while ConvNeXt-Tiny provides a competitive convolutional baseline. In contrast, VGG-16/19 underperform despite larger parameter counts, indicating that modern ConvNet/Transformer families better capture the subtle motif and brushwork cues present in Matryoshka imagery. Confusion matrices (Figure 10) show that the most consequential error mode remains *non-RU predicted as RU*, motivating complementary evidence beyond RGB appearance.

We evaluate whether LLM-generated captions alone contain sufficient information for style and authenticity prediction. Table 7 reports test macro-F1 on Qwen3 captions across four text encoders. Compact sentence-transformer embeddings (MiniLM) provide the strongest and most stable performance, while larger instruction-tuned encoders are unstable under limited supervision. Despite improved descriptive specificity with Qwen3 (Table 5), text-only models still exhibit dangerous authenticity confusions (non-RU $\rightarrow$ RU), consistent with caption priors that over-ascribe "traditional Russian" semantics to visually plausible replicas.

We next fuse visual embeddings from a 2D backbone with caption embeddings to mitigate text priors and improve decision confidence. Table 8 reports ConvNeXt-Tiny results for early, mid, and late fusion. Mid fusion achieves the lowest validation loss and strong test accuracy on the 8-way task, while late fusion yields the best test authenticity accuracy, making it the most robust default configuration under risk-sensitive authentication requirements. Training curves (Figures 14, 16) and fused latent-space projections (Figure 15) indicate stable optimization and improved separability for authenticity labels relative to unimodal baselines.

To validate geometry as an explicit evidence stream, we extract BMA skeletons from ROI-cropped silhouettes (Figure 21) and train a lightweight convolutional autoencoder to reconstruct dense silhouettes from sparse skeleton inputs. Reconstruction metrics (Figures 23, 24) saturate rapidly, and qualitative reconstructions (Figure 25) show reliable recovery of global Matryoshka shape (base expansion, shoulder transitions, and body curvature). These results establish skeletons as a compact, learnable geometric compression modality suitable for downstream reconstruction and fusion.

We benchmark two 3D mesh generation variants: (A) reconstruction from ground-truth silhouettes and (B) reconstruction from Skel-AE predicted silhouettes. Table 12 shows that pipeline (B) achieves near-lossless parity with (A), with Avg IoU $= 0.979$ and Avg Dice $= 0.989$, while slightly reducing generation time. This supports using Skel-AE as a practical surrogate when ground-truth masks are unavailable or withheld for end-to-end evaluation. We then evaluate point-cloud backbones on reconstructed geometry under late fusion with Qwen embeddings (Table 13). Across models, authenticity is consistently more learnable than fine-grained style from geometry alone, reflecting that provenance cues correlate with robust global shape features that survive reconstruction noise.

Finally, we report a geometry-only zero-shot completion baseline that expands partial point clouds using depth projection and back-projection (Figure 36). The method produces stable, contiguous completions consistent with the visible surface but does not incorporate category priors (e.g., axial symmetry), resulting in envelope-like extrapolations. This provides a data-agnostic reference point

 doi:10.20944/preprints202602.0036.v1

that motivates diffusion-based 3D priors (e.g., Hunyuan3D) for improved semantic plausibility and global consistency.

6. Discussion

The core outcome of this work is not limited to Matryoshka dolls; it is a general, auditable blueprint for *evidence-driven authentication* under limited ground truth. The pipeline decomposes the decision into complementary evidence streams—appearance (2D), semantics (text), and geometry (BMA/3D)—that can be re-instantiated for many cultural artifacts where provenance is uncertain and labels are expensive. **Transferable components include:** (i) **Controlled capture + QC-first dataset design**, which is applicable to small glossy objects (ceramics, lacquerware, carved wood, coins) where lighting and glare dominate variance; (ii) **Caption-as-modality construction** using open VLMs to generate scalable "image reports" that can be searched, embedded, and fused with vision features when human annotation is limited; (iii) **Skeletonization as geometric compression**, which reduces texture dependence and provides a compact representation that is recoverable by lightweight decoders, enabling downstream shape reasoning even when full photogrammetry is infeasible; (iv) **Surrogate reconstruction benchmarking** (GT masks vs. predicted masks), which is broadly useful for pipelines that need to generate geometry at scale without requiring dense manual segmentation.

Across unimodal and multimodal experiments, the dominant operational risk is consistent: *visually plausible replicas are frequently mistaken for authentic items* in RGB-only and text-only settings. The results suggest that (a) captions improve context but preserve semantic priors, and (b) complementary evidence streams are essential for verification-grade decisions. A second lesson is methodological: the reconstruction benchmark demonstrates that learned silhouette surrogates can achieve near-lossless parity with ground-truth pipelines, enabling larger ablations and longer-run 3D experiments without prohibitive manual labeling.

This study contributes novelty along three axes: (1) a **private-collection, controlled multi-view dataset** designed explicitly for both fine-grained style recognition and coarse authenticity assessment with physical cue logging; (2) a **BMA-centered geometric modality** used not only for visualization but as a compression-and-reconstruction mechanism that supports downstream 3D generation and learning; (3) an **auditable multimodal authentication framing** that treats geometry and language as explicit evidence streams rather than auxiliary features, enabling interpretation (e.g., Grad-CAM, latent structure) and risk-aware reasoning.

6.1. Limitations and implications

The present study has several limitations that motivate future work. First, authenticity labeling for private collections is inherently uncertain; formalizing an annotation protocol and reporting adjudication statistics would improve reproducibility. Second, fine-grained style discrimination from reconstructed geometry remains challenging, suggesting that higher-fidelity 3D capture (or explicit texture + inscription integration) is needed for large taxonomies. Third, generative priors for 3D completion introduce stochasticity; therefore, completion should be treated as probabilistic evidence and paired with calibration and uncertainty reporting to avoid overconfident decisions in deployment.

7. Conclusion and Future Work

In this work, we introduced an end-to-end 2D–3D–Text pipeline for the analysis and authentication of Matryoshka (nesting) dolls, grounded in a new, fully rights–clear dataset comprising 168 turntable videos and 27,387 labeled frames spanning eight semantic classes and three authenticity labels. The framework unifies strong 2D visual backbones, BMA-based skeletonization with convolutional autoencoding, 3D reconstruction via surface-of-revolution lifting and Poisson meshing, and multiple text / generative-AI components into a single, cultural-heritage–oriented system.

Answer to RQ1 – 2D multi-task baselines. Our first question asked whether a carefully curated Matryoshka dataset can support strong 2D baselines for both semantic and authenticity prediction, and

 doi:10.20944/preprints202602.0036.v1

which architectures are most suitable. Experiments with five backbones (ConvNeXt-Tiny, VGG-16 BN, VGG-19 BN, Swin-Tiny, and ViT-Base) show that transformer-style encoders consistently outperform classical VGG-style CNNs on this benchmark. Swin-Tiny achieves the best overall performance, with test accuracies of 82.98% for the 8-class category task and 85.08% for the 3-way authenticity task, and the highest macro-AUPRC across both tasks. ConvNeXt-Tiny is a strong CNN baseline but generalizes less reliably on authenticity, while VGG-16 and VGG-19 underperform despite their larger parameter counts. Confusion-matrix analysis further indicates that the most consequential residual error mode is *non_authentic / non-RU/replica* being misclassified as *russian_authentic*, whereas the inverse error (authentic $\rightarrow$ replica) occurs less frequently. Grad-CAM visualizations suggest that 2D networks attend to semantically meaningful painted regions (e.g., faces, kokoshnik headpieces, and central apron motifs) but remain largely insensitive to subtle geometric cues (e.g., shell fit, lathe signatures) that are often decisive for authentication. Collectively, these findings confirm that the dataset supports strong 2D baselines and nuanced stylistic analysis, while also highlighting that 2D appearance alone is insufficient for high-confidence authentication.

Answer to RQ2 – BMA skeletons as compressed geometry. The second question concerned whether BMA skeletonization and a convolutional autoencoder (CA) can serve as a compact yet faithful geometric representation. Our BMA pipeline extracts medial skeletons from binarized silhouettes derived from video frames, and the CA learns to reconstruct dense masks from these extremely sparse inputs. Training curves show fast, stable convergence of the binary cross-entropy objective, with validation loss stabilizing around 0.047 and strong visual overlap between predicted and ground-truth silhouettes. Qualitatively, the autoencoder reliably "inflates" 1-pixel-wide skeletons into volumetric shapes that preserve the characteristic Matryoshka profile (bulbous base and smaller head), while smoothing only very fine boundary details. PCA projections of the 64-dimensional latent space reveal structured manifolds aligned with doll identity and viewing angle, indicating that the latent codes capture meaningful geometric variation rather than a degenerate compression [29]. These results support BMA skeletons and learned latents as an effective geometric bottleneck: they substantially reduce storage while retaining sufficient information for accurate silhouette reconstruction and downstream 3D processing. Building on this representation, our 3D stage generates axially symmetric point clouds via surface-of-revolution lifting and produces watertight meshes via Poisson reconstruction. We further show that complete Matryoshka families can be normalized and nested within a shared 3D coordinate system without intersections. Overall, RQ2 is answered in the affirmative, and the BMA-based latent space is validated as a practical bridge from 2D imagery to 3D shape analysis.

Answer to RQ3 – Foundation models, bias, and failure modes. Our third question investigated how existing foundation models perceive and generate Matryoshkas. LLM-guided prompt templates designed for each class (Artistic, Drafted, Merchandise, Non-Authentic, Non-Matryoshka, Political, Religious, Russian Authentic) successfully steer both proprietary image models and Stable Diffusion toward plausible class-conditioned samples, producing diverse grids that reflect the intended high-level styles. In contrast, experiments with the open-source ViT-GPT2 captioner reveal a systematic semantic mismatch: rather than identifying Matryoshkas, the model repeatedly describes scenes using common household-object concepts (e.g., *vases*, *flowers*, *toilets*, *stuffed animals*), consistent with the observed word-cloud and caption statistics. This failure mode suggests strong pretraining-distribution bias and insufficient grounding in the Matryoshka domain. Consequently, such captions are unsuitable as supervisory signals for authentication and would likely contaminate a text modality if used naively. By comparison, curated Gemini-based descriptions (explicitly prompted to mention style, school, and authenticity) provide semantically rich but currently small-scale text data. Overall, RQ3 is answered conditionally: foundation models can be useful as assisted generators and annotators, but they must be treated as biased, imperfect instruments that require explicit auditing, filtering, and/or domain adaptation for cultural-heritage use.

Answer to RQ4 – Transfer of zero-shot 3D completion priors to axially symmetric cultural artifacts. Our fourth question asked how well a zero-shot point-cloud completion pipeline, trained

 doi:10.20944/preprints202602.0036.v1

on everyday objects, transfers to axially symmetric cultural artifacts such as Matryoshka dolls. The results indicate that the completion prior transfers in a limited but encouraging way: from partial observations and occlusions, the pipeline often recovers coherent global structure and produces stable, visually consistent completions. However, the geometry-only setting prioritizes surface continuity and visibility over object-specific semantics, and it does not reliably recover fine-grained authenticity-relevant details. As a consequence, the current zero-shot completion stage is best interpreted as a data-agnostic baseline for 3D reconstruction quality rather than a complete solution for authentication. These findings motivate integrating semantic 3D generative priors (e.g., 3D diffusion) and explicit artifact-aware constraints to improve fidelity while preserving geometric consistency.

Future work. Several directions follow directly from the present study:

- *Full 3D backbones and multimodal fusion.* We will train dedicated 3D backbones (e.g., PointNet++, DGCNN, point transformers, and mesh-based networks such as MeshCNN) on reconstructed meshes and nested families, and integrate their features with ConvNeXt/Swin embeddings via calibrated late fusion (e.g., an MLP with temperature scaling) to improve confidence calibration and reduce high-impact authenticity errors.
- *Richer and cleaner text streams.* We will expand curated LLM-based descriptions (including multilingual prompts), explore retrieval-augmented text encoders, and explicitly exclude off-the-shelf caption streams unless they are filtered, audited, and/or retrained for Matryoshka-specific grounding.
- *Larger datasets and expert labels.* Extending the dataset with museum-grade collections and expert-provided regional/era annotations would enable finer-grained school classification, stronger replica evaluation protocols, and more rigorous calibration and uncertainty studies.
- *From geometric completion to semantic 3D diffusion.* We will augment the current geometric zero-shot completion stage with 3D diffusion (or hybrid score-based) models to increase reconstruction fidelity, aiming to preserve low Chamfer distance while recovering fine-grained, semantically meaningful details relevant to authenticity.
- *Deployment, interaction, and ethics.* Finally, we plan to operationalize the pipeline as an interactive web and AR/VR tool for collectors, curators, and educators, with explicit uncertainty reporting, human-in-the-loop review, and a dedicated study of the ethical implications and failure costs of automated cultural-heritage authentication.

In summary, this paper establishes a first end-to-end benchmark for Matryoshka authentication using calibrated 2D–3D–Text fusion, identifies key limitations of 2D-only and foundation-model–only approaches (including domain bias in captioning), and lays the groundwork for future multimodal systems that can more reliably protect and interpret cultural heritage through nested geometric and semantic representations.

Author Contributions: Conceptualization, Y.K. and S.S; methodology, Y.K. and S.S.; software, Y.K. and S.S.; validation, Y.K.; formal analysis, Y.K. and S.S.; investigation, Y.K. and S.S.; resources, Y.K.; data curation, Y.K.; writing—original draft preparation, Y.K. and S.S.; writing—review and editing, Y.K.; visualization, Y.K. and S.S.; supervision, Y.K.; project administration, Y.K.; funding acquisition, Y.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. Publication is supported by Kean University (Union, NJ).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data will be released to a public archive upon the paper acceptance.

Acknowledgments: The authors gratefully acknowledge the financial support Kean University.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

 doi:10.20944/preprints202602.0036.v1

Abbrev.	Definition
AE	Autoencoder
AUPRC	Area Under the Precision–Recall Curve
BCE	Binary Cross-Entropy
BMA	Blum Medial Axis
CA	Convolutional Autoencoder
CE	Cross-Entropy
CNN	Convolutional Neural Network
COLMAP	Structure-from-Motion / Multi-View Stereo pipeline (software)
ControlNet	Conditioning network for diffusion models
Dice	Dice Similarity Coefficient
DGCNN	Dynamic Graph CNN
FPS	Frames per Second
GD	Gradient Descent
Grad-CAM	Gradient-weighted Class Activation Mapping
HBIM	Heritage Building Information Modeling
ICP	Iterative Closest Point
IoU	Intersection over Union
LLM	Large Language Model
MML	Multimodal Machine Learning
MND	Matryoshka Nesting Dolls
MVS	Multi-View Stereo
OCR	Optical Character Recognition
PCA	Principal Component Analysis
PCT	Point Cloud Transformer
ROI	Region of Interest
SfM	Structure-from-Motion
SD	Stable Diffusion
SfM/MVS	Structure-from-Motion / Multi-View Stereo
t-SNE	t-distributed Stochastic Neighbor Embedding
UCD	Unidirectional Chamfer Distance
ViT	Vision Transformer
VLM	Vision–Language Model
YOLO	You Only Look Once

References

1. F. Remondino and S. Campana, "Heritage Recording and 3D Modeling with Photogrammetry," *Remote Sensing*, vol. 3, no. 6, pp. 1104–1138, 2011. Available: MDPI. https://www.mdpi.com/2072-4292/3/6/1104
2. J. L. Schönberger and J.-M. Frahm, "Structure-from-Motion Revisited," in *CVPR*, 2016, pp. 4104–4113. https://openaccess.thecvf.com/content_cvpr_2016/html/Schonberger_Structure-From-Motion_Revisited_CVPR_2016_paper.html
3. J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise View Selection for Unstructured Multi-View Stereo," in *ECCV*, 2016, pp. 501–518. https://demuc.de/papers/schoenberger2016mvs.pdf
4. P. Cignoni *et al.*, "MeshLab: An Open-Source Mesh Processing Tool," *Eurographics Italian Chapter Conf.*, 2008. https://diglib.eg.org/items/cafa9acd-c34e-4247-be5c-e4b7beb14a46
5. R. Pintus, S. M. P. Gallo, and M. Callieri, "A Survey of Geometric Analysis in Cultural Heritage," *Computer Graphics Forum*, 2016. https://dl.acm.org/doi/10.1111/cgf.12668
6. A. Dosovitskiy *et al.*, "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," 2020. https://arxiv.org/abs/2010.11929
7. Z. Liu *et al.*, "A ConvNet for the 2020s," in *CVPR*, 2022. https://openaccess.thecvf.com/content/CVPR2022/papers/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.pdf
8. C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," in *CVPR*, 2017. https://openaccess.thecvf.com/content_cvpr_2017/papers/Qi_PointNet_Deep_Learning_CVPR_2017_paper.pdf

 doi:10.20944/preprints202602.0036.v1

9. C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," in *NeurIPS*, 2017. https://arxiv.org/abs/1706.02413
10. R. Hanocka *et al.*, "MeshCNN: A Network with an Edge," *ACM TOG (SIGGRAPH)*, 2019. https://arxiv.org/abs/1809.05910
11. A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *ICML*, 2021. https://arxiv.org/abs/2103.00020
12. M. Li *et al.*, "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models," 2021. https://arxiv.org/abs/2109.10282
13. Z. Kuang *et al.*, "MMOCR: A Comprehensive Toolbox for Text Detection, Recognition and Understanding," *ACM MM*, 2021. https://dl.acm.org/doi/10.1145/3474085.3478328
14. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *ICML*, 2017. https://proceedings.mlr.press/v70/guo17a.html
15. R. R. Selvaraju *et al.*, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in *ICCV*, 2017. https://openaccess.thecvf.com/content_iccv_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.pdf
16. T. Zheng, C. Zhe, and J. Yuan, "PointCloud Saliency Maps," in *ICCV*, 2019.
17. G. Earl, K. Martinez, and T. Malzbender, "Archaeological Applications of Polynomial Texture Mapping," *Journal of Archaeological Science*, 2010.
18. T. Malzbender, D. Gelb, and H. Wolters, "Polynomial Texture Maps," in *SIGGRAPH*, 2001. https://www.hpl.hp.com/techreports/2001/HPL-2001-33.html
19. R. Wightman. PyTorch Image Models. GitHub repository, 2019. Available at: https://github.com/huggingface/pytorch-image-models. doi:10.5281/zenodo.4414861.
20. K. Simonyan and A. Zisserman. Very deep CNNs for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
21. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. CoRR, abs/2103.14030, 2021. Available at: https://arxiv.org/abs/2103.14030.
22. R. Wightman. PyTorch Image Models. GitHub repository, 2019. Available at: https://github.com/huggingface/pytorch-image-models. doi:10.5281/zenodo.4414861.
23. TorchVision. TorchVision models — torchvision 0.8.1 documentation. PyTorch, 2020. Available at: https://docs.pytorch.org/vision/0.8/models.html.
24. G. Wong, Y. Kumar, J. J. Li, and D. Kruger. Real-time object detection and skeletonization for motion prediction in video streaming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 28, pp. 29712–29714, 2025.
25. Y. Kumar, Z. Gordon, O. Alabi, J. Li, K. Leonard, L. Ness, and P. Morreale. ChatGPT translation of program code for image sketch abstraction. *Applied Sciences*, vol. 14, no. 3, p. 992, 2024.
26. K. Leonard, G. Morin, S. Hahmann, and A. Carlier. A 2D shape structure for decomposition and part similarity. In *Proceedings of the 23rd International Conference on Pattern Recognition (ICPR)*, Cancun, Mexico, 4–8 December 2016, pp. 3216–3221.
27. Y. Kumar, K. Huang, Z. Gordon, L. Castro, E. Okumu, P. Morreale, and J. J. Li. Transformers and LLMs as the new benchmark in early cancer detection. In *ITM Web of Conferences*, vol. 60, p. 00004, 2024. latex
28. Y. Kumar, K. Huang, C.-C. Lin, A. Watson, J. J. Li, P. Morreale, and J. Delgado. Applying Swin architecture to diverse sign language datasets. *Electronics*, vol. 13, no. 8, p. 1509, 2024.
29. D. Kruger, Y. Kumar, and J. J. Li, "Parametric Matching for Improved Data Compression," in *Proc. 2025 Data Compression Conference (DCC)*, IEEE, 2025, pp. 383–383.
30. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Björn, "High-Resolution Image Synthesis With Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10684–10695.
31. P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution," *arXiv preprint* arXiv:2409.12191, 2024.
32. J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond," *arXiv preprint* arXiv:2308.12966, 2023.

 doi:10.20944/preprints202602.0036.v1

33. S. Choi, Q.-Y. Zhou, S. Miller, and V. Koltun, "A large dataset of object scans," *arXiv preprint arXiv:1602.02481*, 2016.
34. RunDiffusion, Juggernaut XL v9 + RunDiffusion Photo v2 Official," *Hugging Face*, 2024. Available: https://huggingface.co/RunDiffusion/Juggernaut-XL-v9. Accessed: 6 Dec. 2025.
35. K. Chedraoui, *Nano Banana Pro Is the Best AI Image Tool I've Tested. It's Also Deeply Troubling*. CNET, Dec. 4, 2025. [Online]. Available: https://www.cnet.com/tech/services-and-software/google-nano-banana-pro-ai-image-generator-review/. [Accessed: Dec. 2025].
36. C. Lin, L. Liu, C. Li, L. Kobbelt, B. Wang, S. Xin, and W. Wang, "SEG-MAT: 3D Shape Segmentation Using MA Transform," *arXiv preprint arXiv:2010.11488*, 2020.
37. R. Pierdicca, M. Paolanti, F. Matrone, M. Martini, C. Morbidoni, E. S. Malinverni, E. Frontoni, and A. M. Lingua, "Point Cloud Semantic Segmentation Using a Deep Learning Framework for Cultural Heritage," *Remote Sensing*, vol. 12, no. 6, p. 1005, 2020.
38. X. Yang, X. Bai, D. Yu, and L. J. Latecki, "Shape Classification Based on Skeleton Path Similarity," in *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, LNCS, vol. 4679, pp. 375–386, Springer, 2007.
39. X. Zeng, A. Vahdat, F. Williams, Z. Gojcic, O. Litany, S. Fidler, and K. Kreis, "LION: Latent Point Diffusion Models for 3D Shape Generation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
40. Y. Liu, N. Pears, P. L. Rosin, and P. Huber (eds.), *3D Imaging, Analysis and Applications*, 2nd ed. Cham, Switzerland: Springer, 2020.
41. Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, "SkeletonNet: Mining Deep Part Features for 3D Action Recognition," *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 731–735, 2017.
42. X. Zhu, R. Zhang, B. He, Z. Zeng, S. Zhang, and P. Gao, "PointCLIP V2: Adapting CLIP for Powerful 3D Open-World Learning," *arXiv preprint arXiv:2211.11682*, 2022.
43. A. Sanghi, H. Chu, J. G. Lambourne, Y. Wang, C.-Y. Cheng, M. Fumero, and K. R. Malekshan, "CLIP-Forge: Towards Zero-Shot Text-to-Shape Generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18603–18613, 2022.
44. B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "DreamFusion: Text-to-3D Using 2D Diffusion," *arXiv preprint arXiv:2209.14988*, 2022.
45. Make Sense AI, *makesense.ai: Free Online Tool for Image Annotation and Labeling*, Accessed: 2025-12-13. Available: `https://www.makesense.ai/`
46. S. Choi, Q.-Y. Zhou, S. Miller, and V. Koltun, "A Large Dataset of Object Scans," Technical Report, arXiv:1602.02481, 2016.
47. Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A Modern Library for 3D Data Processing," arXiv:1801.09847, 2018.
48. W. Yuan, T. Khot, D. Held, C. Mertz, and M. Hebert, "PCN: Point Completion Network," in *Proceedings of the International Conference on 3D Vision (3DV)*, 2018, pp. 728–737.
49. P. Xiang, X. Wen, Y.-S. Liu, Y.-P. Cao, P. Wan, W. Zheng, and Z. Han, "SnowflakeNet: Point Cloud Completion by Snowflake Point Deconvolution with Skip-Transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
50. X. Yu, Y. Rao, Z. Wang, Z. Liu, J. Lu, and J. Zhou, "PoinTr: Diverse Point Cloud Completion with Geometry-Aware Transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
51. A. Li, Z. Zhu, and M. Wei, "GenPC: Zero-shot Point Cloud Completion via 3D Generative Priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
52. P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
53. Tencent Hunyuan3D Team, "Hunyuan3D 2.0: Scaling Diffusion Models for High-Resolution Textured 3D Assets Generation," *arXiv preprint arXiv:2501.12202*, 2025.
54. T. Huang, Z. Yan, Y. Zhao, and G. H. Lee, "ComPC: Completing a 3D Point Cloud with 2D Diffusion Priors," in *International Conference on Learning Representations (ICLR)*, 2025.
55. B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," *arXiv preprint arXiv:2003.08934*, 2020.
56. B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, 2023.

 doi:10.20944/preprints202602.0036.v1

57. R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot One Image to 3D Object," *arXiv preprint arXiv:2303.11328*, 2023.
58. B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "DreamFusion: Text-to-3D Using 2D Diffusion," *arXiv preprint arXiv:2209.14988*, 2022.
59. Y. Kasten, O. Rahamim, and G. Chechik, "Point Cloud Completion with Pretrained Text-to-Image Diffusion Models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2024.
60. J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
61. J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.
62. W. Peebles and S. Xie, "Scalable Diffusion Models with Transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4195–4205.
63. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695.
64. P. J. Besl and N. D. McKay, "A Method for Registration of 3-D Shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.
65. A. X. Chang *et al.*, "ShapeNet: An Information-Rich 3D Model Repository," *arXiv preprint arXiv:1512.03012*, 2015.
66. S. Sengupta and Y. Zhou, "Zero-Shot Point Cloud Completion via 3D Diffusion Priors," unpublished manuscript, 2025.
67. Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A Modern Library for 3D Data Processing," *arXiv preprint arXiv:1801.09847*, 2018.