## A. Generative Potential of Real-IAD-MVN

To showcase the richness of the learned representation in our new dataset and to explore future research directions, we present an **Awareness-Guided Generative Module**. This module is designed to validate the deep, intrinsic understanding of 3D structure that can be learned from our high-fidelity 2D (RGB) and Pseudo-3D (Normal Vector) data.

Its purpose is to generate a high-quality geometric representation (e.g., a single-channel depth image) using only the RGB and normal vector (NV) inputs provided by our Real-IAD-MVN dataset. The success of this complex cross-modal generation task serves as compelling evidence that our new data modality (MV-RGB + MV-NV) captures a profound understanding of object structure, potentially enabling hardware-agnostic applications where expensive 3D scanners can be replaced.

Concretely, the module follows a two-stage pipeline conditioned by a hierarchical fusion mechanism. First, RGB and NV features are encoded and fused by a **Hierarchical Dual Conditioner (HDC)**. The HDC performs multi-level cross-modal interaction and injects the resulting conditional features into a **Denoising Unet**, which predicts a latent geometric representation through a conditional diffusion process. Second, the denoised latent code is forwarded to an **Adversarial Dense Generation** decoder that reconstructs a high-fidelity depth image. The overall design therefore combines cross-modal prototype-aware conditioning, diffusion-based latent generation, and adversarial dense decoding in a unified generation framework.

### A.1. Two-Stage Generation for High-Fidelity Depth

We first generate a latent geometric representation using diffusion, then decode it into a dense depth image using an adversarial network.

**1. Diffusion-based Latent Generation.** We adopt a conditional denoising diffusion probabilistic model (DDPM) to generate a latent representation of the geometry. A forward process gradually adds Gaussian noise to a ground truth latent vector, and a **Denoising Unet** is trained to reverse this process. The reverse process is conditioned on the fused features from the RGB image ($I$) and the pseudo-3D normal vectors ($P$):

$$p_\theta(x_{t-1}|x_t, I, P) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, I, P), \sigma_t^2 \mathbf{I}) \quad \text{(A1)}$$

The Unet, $\epsilon_\theta$, is trained to predict the noise added at each timestep $t$:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t, I, P)\|_2^2\right] \quad \text{(A2)}$$

**2. Adversarial Dense Generation.** The denoised latent representation from the Unet is passed to a final **Decoder** network, which generates the high-resolution, single-channel depth image $\hat{I}_{\text{depth}}$. This stage is trained with a combination of a reconstruction loss and an adversarial loss to ensure high fidelity and realism:

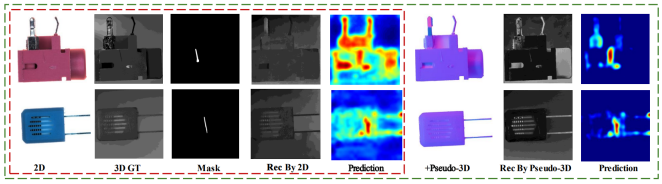


Figure A1. Qualitative comparison of anomaly segmentation from the generative module's output on the Real-IAD D³ dataset. The right panel (+Pseudo-3D) shows that generating geometry from normal vectors (Rec By Pseudo-3D) leads to a sharper, more accurate anomaly prediction compared to the 2D-only baseline (left panel).

$$\mathcal{L}_{\text{reconstruct}} = \mathbb{E}\left[\|\hat{I}_{\text{depth}} - I_{\text{depth\_GT}}\|_2^2\right] \quad \text{(A3)}$$

$$\mathcal{L}_{\text{adv}} = \mathbb{E}\left[\log(1 - D(\hat{I}_{\text{depth}}))\right] + \mathbb{E}\left[\log D(I_{\text{depth\_GT}})\right] \quad \text{(A4)}$$

where $D$ is a discriminator network distinguishing generated images from ground truth (GT) depth images.

### A.2. Hierarchical Conditioning for Input Fusion

The conditioning of the Denoising Unet is orchestrated by a **Hierarchical Dual Conditioner (HDC)**, which effectively fuses the image and normal vector modalities. It uses resolution-aware information routing, extracting features $F_{\mathcal{M}}^l$ at $L$ hierarchical levels. The core is a Bi-Modal Cross-Attention (BCA) module that uses features from the Unified Cross-Modal Prototype Extraction (UCP) module (detailed in our main paper's baseline) as a rich feature source.

$$\text{CA}_{\mathcal{M}}(Q, \tilde{F}_{\mathcal{M}}^l) = \text{Softmax}\left(\frac{QW_Q^{\mathcal{M}}(W_K^{\mathcal{M}}\tilde{F}_{\mathcal{M}}^l)^T}{\sqrt{d_k}}\right) W_V^{\mathcal{M}}\tilde{F}_{\mathcal{M}}^l \quad \text{(A5)}$$

To dynamically balance the modalities, a weighting coefficient $\alpha_l$ is computed for each level $l$, based on the estimated mutual information $\mathcal{I}$ that each modality's features provide with respect to the Unet's query features $Q$:

$$\alpha_l = \sigma\left(\beta_l \cdot \frac{\mathcal{I}(Q; \tilde{F}_P^l)}{\mathcal{I}(Q; \tilde{F}_P^l) + \mathcal{I}(Q; \tilde{F}_I^l)}\right) \quad \text{(A6)}$$

The final conditional features injected into the Unet are a weighted sum of the attention outputs, ensuring that the most relevant information from each modality (RGB and NV) is used at the most appropriate stage of generation.

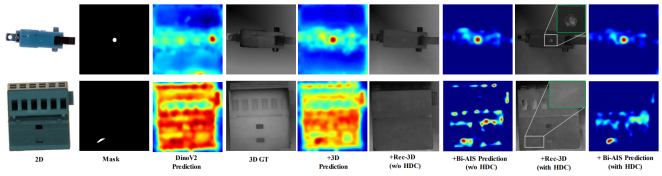


Figure A2. Qualitative results of the generative module's ablation study on Real-IAD D³. The baseline 2D prediction (DinoV2) is diffuse. Adding 3D reconstruction (+Rec-3D) improves focus. The full model (+Bi-AIS Prediction (with HDC)) demonstrates the sharpest and most accurate localization by effectively fusing the RGB and normal vector data.

### A.3. Qualitative Analysis of Generative Module

While the quantitative results of this generative model are reserved for future work, we provide qualitative results here to demonstrate its effectiveness, which further validates the quality of our Real-IAD-MVN data.

The qualitative results in Figure A1 (conducted on the similar Real-IAD D³ dataset) visually demonstrate the principle of defect enhancement. The left panel shows the prediction from a 2D-only baseline, which produces a diffuse and inaccurate anomaly map. In contrast, the right panel shows that after generating a geometric representation guided by pseudo-3D information (normal vectors), the model's reconstruction is sharper, and the final prediction is precisely focused on the true anomalous regions (e.g., the bent pins), providing a much clearer signal for detection.

The progressive improvement in localization accuracy is also visualized in the ablation study in Figure A2. The baseline model's (DinoV2) heatmap is noisy and unfocused. Adding the reconstruction module with pseudo-3D data (+Rec-3D) begins to consolidate the activation. Finally, the full model with the Hierarchical Dual Conditioner (+Bi-AIS Prediction (with HDC)) produces a remarkably clean and accurate anomaly map, demonstrating the critical role of sophisticated fusion of the RGB and normal vector data.

## B. Real-IAD-MVN Dataset Visualizations

To provide a comprehensive overview of the data included in Real-IAD-MVN, Figure A3 visualizes all 20 object categories. For each category, we display the anomaly-free top-down view, followed by all five distinct, calibrated viewpoints (Side View 0 through Side View 4). Each viewpoint is represented by its (1) RGB image, (2) high-fidelity PS Normal Vector (NV) map, and (3) the corresponding ground truth anomaly mask (shown in white, if an anomaly is present in that view).

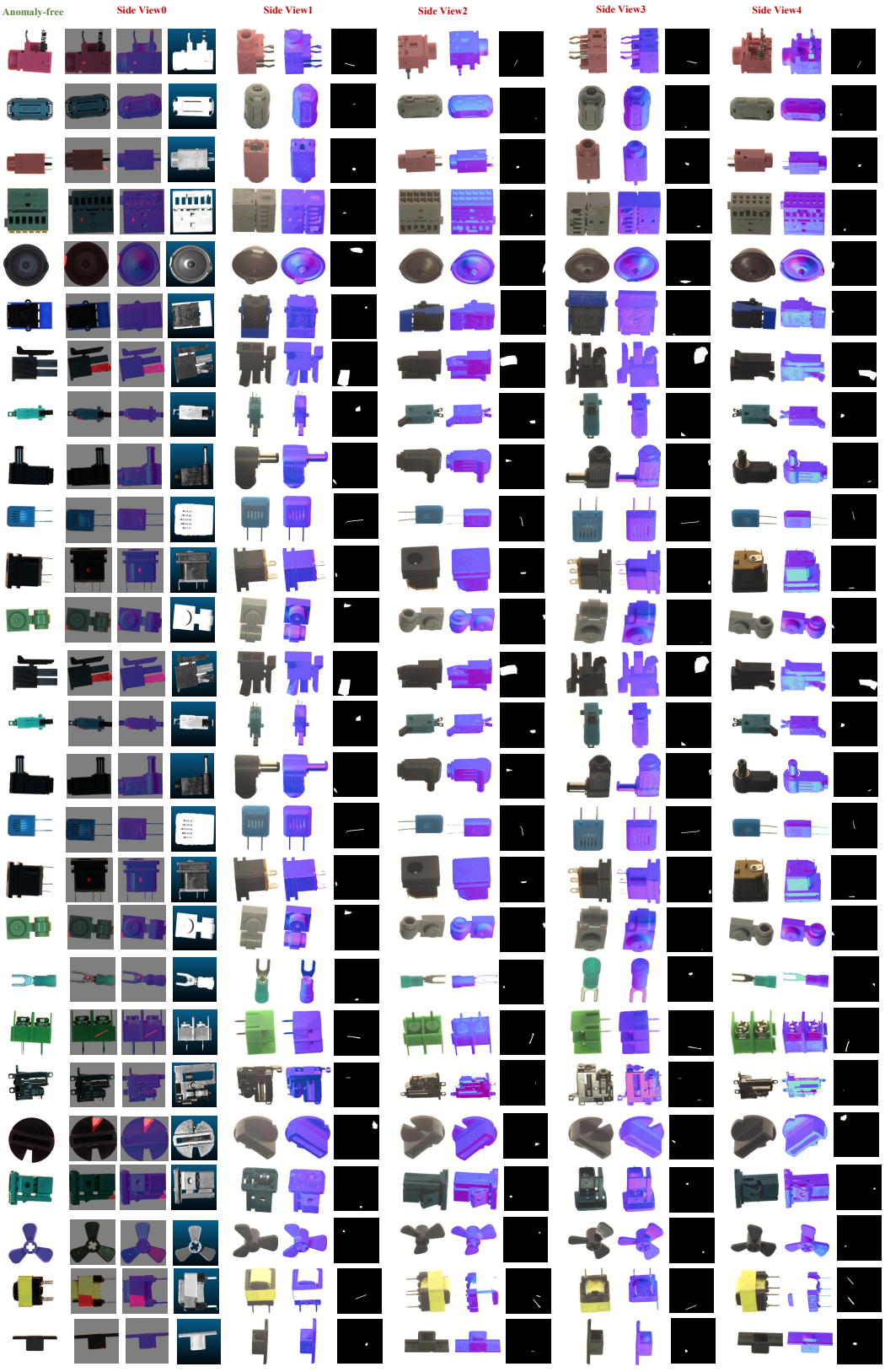


Figure A3. Comprehensive visualization of all 20 categories in the Real-IAD-MVN dataset. Each row corresponds to one object category. The first column shows the anomaly-free top-down RGB image. The subsequent columns (Side View 0-4) display the five captured viewpoints. For each viewpoint, we show the RGB image, its corresponding high-fidelity PS Normal Map (NV), and the pixel-level ground truth anomaly mask.

## C. Camera Calibration Parameters

In Section 3.1 of our main paper, we describe our integrated multi-view photometric stereo (MVPS) gantry, which operates with five pre-calibrated viewpoints. To enable and encourage future research in multi-view 3D reconstruction, geometric registration, and novel view synthesis (e.g., NeRF-based tasks), we provide the full calibration parameters for four of our primary acquisition cameras below. The parameters are provided in the '.yml' format, compatible with standard calibration tools.

### C.1. Camera 1 Parameters

```
calibration_file_ver: 3.1
calibration_ver: 3.1.359264
calibration_time: 2025-1-24 15:6:57
calibration_type: MVP_CALIB_TYPE_NORMAL
frame_num: 1
image_width: 4096
image_height: 3000
board_width: 0
board_height: 0
board_dx: 2.0000000000000000e+000
board_dy: 2.0000000000000000e+000
points_num: 0
H_matrix:
  rows: 3
  cols: 3
  data:
    - 7.5142908289968972e+001
    - 4.9162084131020807e+000
    - 8.2109891582446537e+001
    - -1.7980084333606616e+000
    - 5.7724585197937522e+001
    - 1.0859333933073324e+003
    - 8.2071064611873192e-005
    - 2.3720123199760068e-003
    - 1.0000000000000000e+000
distortion_model: 3
distortion_level: 2
distortion_matrix:
  rows : 12
  cols : 1
  data:
    - -7.0994615634284253e-001
    - 1.5565761184853105e+001
    - -6.7707702880165407e-002
    - -1.2152307380140845e-002
    - -1.1598420108419958e+002
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
intrinsic_matrix:
  rows : 3
  cols : 3
  data:
    - 1.9426147574516781e+004
    - 0.0000000000000000e+000
    - 1.9428998682433419e+003
    - 0.0000000000000000e+000
    - 2.0040639975377839e+004
    - -1.4103926763176123e+003
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 1.0000000000000000e+000
extrinsic_matrix:
  rows : 3
  cols : 4
  data:
    - 9.9953780819267646e-001
    - 4.1006888450079194e-003
    - -3.0122323024239783e-002
    - -2.4804538528430975e+001
    - -2.1737068720812217e-002
    - 7.8910936074326898e-001
    - -6.1386799609588383e-001
    - 3.2255977192006476e+001
    - 2.1252525421862352e-002
    - 6.1423904234291860e-001
    - 7.8883377781687780e-001
    - 2.5895272008921592e+002
```

### C.2. Camera 2 Parameters

```
calibration_file_ver: 3.1
calibration_ver: 3.1.359264
calibration_time: 2025-1-24 15:4:51
calibration_type: MVP_CALIB_TYPE_NORMAL
frame_num: 1
image_width: 4096
image_height: 3000
board_width: 0
board_height: 0
board_dx: 2.0000000000000000e+000
board_dy: 2.0000000000000000e+000
points_num: 0
H_matrix:
  rows: 3
  cols: 3
  data:
    - 7.4755050461534353e+001
    - 6.2458574231518309e+000
    - 1.0981382636031776e+003
    - -1.6285738267351551e+000
```

```
    - 5.7629797788179566e+001
    - 1.7663434952193629e+002
    - -4.8201648078641062e-004
    - 2.8343758580036344e-003
    - 1.0000000000000000e+000
distortion_model: 3
distortion_level: 2
distortion_matrix:
  rows : 12
  cols : 1
  data:
    - -2.4410037940517855e+000
    - 4.3015853395902596e+001
    - 2.1299813348685793e-002
    - 8.8716663242255092e-002
    - -3.5257553640216827e+002
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
intrinsic_matrix:
  rows : 3
  cols : 3
  data:
    - 1.9798511319027260e+004
    - 0.0000000000000000e+000
    - -4.3049246222583417e+002
    - 0.0000000000000000e+000
    - 2.0271741030735127e+004
    - 2.4744169291864846e+003
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 1.0000000000000000e+000
extrinsic_matrix:
  rows : 3
  cols : 4
  data:
    - 9.9188951598251840e-001
    - 9.9339040733826808e-002
    - 7.9290245743398075e-002
    - 2.0339137531916062e+001
    - -5.6639905656809125e-003
    - 6.5775238666472491e-001
    - -7.5321292942150830e-001
    - -2.9859351241362621e+001
    - -1.2697679825400085e-001
    - 7.4665490879183360e-001
    - 6.5298035183474057e-001
    - 2.6342833350187118e+002
```


### C.3. Camera 3 Parameters


```
calibration_file_ver: 3.1
calibration_ver: 3.1.359264
calibration_time: 2025-1-24 15:54:51
calibration_type: MVP_CALIB_TYPE_NORMAL
frame_num: 1
image_width: 4096
image_height: 3000
board_width: 0
board_height: 0
board_dx: 2.0000000000000000e+000
board_dy: 2.0000000000000000e+000
points_num: 0
H_matrix:
  rows: 3
  cols: 3
  data:
    - 7.5368561341388229e+001
    - 6.6572215304853710e+000
    - 2.2010105906328986e+003
    - 1.2956815846651235e+000
    - 5.6687607049282498e+001
    - 4.9215373641409366e+002
    - 1.3453776666924063e-004
    - 3.0541386844316651e-003
    - 1.0000000000000000e+000
distortion_model: 3
distortion_level: 2
distortion_matrix:
  rows : 12
  cols : 1
  data:
    - 1.8004248377403349e-002
    - -3.0044854730835597e+000
    - -2.5340313073236019e-003
    - 1.5663430597834375e-002
    - 3.9016750900379250e+001
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
intrinsic_matrix:
  rows : 3
  cols : 3
  data:
    - 1.7313015180332859e+004
    - 0.0000000000000000e+000
    - 2.9813759579949819e+003
    - 0.0000000000000000e+000
    - 1.7258484049521408e+004
```

```
    - 1.2130817833681663e+003
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 1.0000000000000000e+000
extrinsic_matrix:
  rows : 3
  cols : 4
  data:
    - 9.9940303465816926e-001
    - -3.2638864426226488e-002
    - -1.1326025119533373e-002
    - -1.0403174684703622e+001
    - 1.5144923326588894e-002
    - 7.0855235884226586e-001
    - -7.0549570237988890e-001
    - -9.6411733094746737e+000
    - 3.1051660398014748e-002
    - 7.0490301411472722e-001
    - 7.0862369074036691e-001
    - 2.3080255579353312e+002
```


### C.4. Camera 4 Parameters


```
calibration_file_ver: 3.1
calibration_ver: 3.1.359264
calibration_time: 2025-1-24 15:56:5
calibration_type: MVP_CALIB_TYPE_NORMAL
frame_num: 1
image_width: 4096
image_height: 3000
board_width: 0
board_height: 0
board_dx: 2.0000000000000000e+000
board_dy: 2.0000000000000000e+000
points_num: 0
H_matrix:
  rows: 3
  cols: 3
  data:
    - 7.2093441984252976e+001
    - 6.3939899920335943e+000
    - 1.2086169768470568e+003
    - 9.8103425121876583e-003
    - 5.3897464619363276e+001
    - 1.2205660967156343e+003
    - 3.1503871509957305e-004
    - 2.7843508343726778e-003
    - 1.0000000000000000e+000
distortion_model: 3
distortion_level: 2
distortion_matrix:
  rows : 12
  cols : 1
  data:
    - 1.8131330524292905e-001
    - -1.2469765989270881e+001
    - 6.3753197148693971e-003
    - 9.0223021157360868e-003
    - 1.5616130476247724e+002
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
intrinsic_matrix:
  rows : 3
  cols : 3
  data:
    - 1.7832381835705353e+004
    - 0.0000000000000000e+000
    - 3.6234241894269994e+003
    - 0.0000000000000000e+000
    - 1.7843555944438864e+004
    - 1.0788558390330131e+003
    - 0.0000000000000000e+000
    - 0.0000000000000000e+000
    - 1.0000000000000000e+000
extrinsic_matrix:
  rows : 3
  cols : 4
  data:
    - 9.9686931222851527e-001
    - -5.1912991967135176e-002
    - -5.9637367497796648e-002
    - -3.3927864116395583e+001
    - -4.6345694251529318e-003
    - 7.1460287931792632e-001
    - -6.9951500744213679e-001
    - 1.9897718916501284e+000
    - 7.8930951491100951e-002
    - 6.9760143788236939e-001
    - 7.1212508645683870e-001
    - 2.5054365609050163e+002
```