The Faculty of Engineering
The Maersk Mc-Kinney Moller Institute
Project Report

Student Name: Mehmet Baha Dursun	**Student ID Number:** 501209

Class and Year:	Master of Science in Engineering Drone Technology and Autonomous Systems 2025-2027
Subject Code and Name:	T550021101-1-F26 Kunstig intelligens værktøjer, (F26)
Lecturer Name(s):	Vinay Chakravarthi Gogineni / Smith Khare / Orkun Furat
Title of Report/Assignment:	Do Self-Supervised Vision Transformers Localize What Surgical Policies Need? Comparing DINOv2 and ResNet Encoders for a dVRK Diffusion Policy
Submission Date:	May 31, 2026

Abstract

The visual encoder is a central design choice when learning visuomotor control for surgical robots, yet it is rarely studied with respect to *what* the encoder localizes and retains. We compare a self-supervised vision transformer (DINOv2 ViT-S/14, at three state-dropout settings) against a convolutional baseline (ResNet34 with spatial-softmax) as the perception front-end of a single, fixed diffusion policy, trained on 84 da-Vinci Research Kit (dVRK) suturing demonstrations (43 890 frames, 10 FPS) in the LeRobot v3 format. Rather than relying on action error alone, we apply a probe suite, decoder feature inversion, a physical pose probe (linear vs. MLP R^2), a modality-attribution score R_{vision}, and a needle-localization patch-similarity query, together with a state-dropout sweep ($p \in \{0, 0.2, 1.0\}$). DINOv2 reaches a lower comparable action validation loss than ResNet34 (0.002 68 vs. 0.004 36); its top-k similarity patches trace the suturing needle across frames, whereas ResNet34 scatters onto surrounding tissue and fails to isolate it. The state-dropout sweep is monotone for DINOv2: R_{vision} rises from 0.76 at p=0 (a residual state-shortcut) to 0.975 at p=0.2 and 0.995 at p=1 (essentially pure vision), removing the shortcut. ResNet34 exposes pose more linearly but misses the task-critical needle. We discuss confounds (resolution, probe sample sizes) and report first-hand closed-loop observations.

Abbreviations and key terms

Term	Meaning
dVRK	da Vinci Research Kit, the open surgical robot platform used here
PSM1 / PSM2	the two Patient-Side Manipulators (robot arms)
DINOv2	a self-supervised Vision Transformer image encoder (Oquab et al., 2024)
ViT-S/14	"Small" Vision Transformer with a 14×14-pixel patch size
ResNet34	a 34-layer supervised convolutional encoder (He et al., 2016)
spatial-softmax	a read-out turning a CNN feature map into 2D keypoint coordinates
DDPM	Denoising Diffusion Probabilistic Model: the noise schedule used in training (Ho et al., 2020)
DDIM	Denoising Diffusion Implicit Model: a faster sampler used at inference (Song et al., 2021)
state dropout (p)	probability of zeroing the proprioceptive state during training
state-shortcut	a policy relying on proprioception instead of vision
R_{vision}	vision share: fraction of the action change caused by the image vs. state
MAE / MSE	mean absolute / squared error
rot6d	6-dimensional continuous rotation representation
EMA	exponential moving average of the model weights

1 Introduction

Autonomous and semi-autonomous assistance in robotic surgery requires policies that map endoscopic images and robot proprioception into fine, safe motions. Diffusion policies (Chi et al., 2023) have become a strong paradigm for such visuomotor control because they model multimodal action distributions and emit smooth action chunks. Every such policy, however, depends on a *visual encoder*, and the encoder determines which parts of the scene the controller can effectively see. For a suturing task this is decisive, because the most task-critical object is a thin, curved needle that occupies only a few pixels and is frequently occluded by tissue and tools.

Convolutional encoders such as ResNet (He et al., 2016), typically combined with a spatial-softmax keypoint read-out (Levine et al., 2016), are the standard front-end in the original diffusion-policy recipe (Chi et al., 2023). Self-supervised vision transformers, in particular DINOv2 (Oquab et al., 2024), offer an alternative whose patch features are known to be semantically structured and to localize objects without labels. The practical question for a surgical deployment is therefore not only "which encoder gives lower action loss?" but "*what does the encoder localize and retain*, and is that what the policy needs?". If a policy can reach low action loss by quietly leaning on proprioception (a *state-shortcut*), it can look accurate offline yet fail when the scene changes.

This report studies that question under one fixed diffusion policy on a real dVRK suturing dataset, varying only the encoder and a single state-dropout regularizer. In the context of the course, the project falls under the Self-Supervised Learning topic (Module 5): as it is carried out individually, we compare two distinct representation-learning methods on one shared dataset, a self-supervised vision transformer (DINOv2) and a supervised convolutional encoder (ResNet34), and benchmark them against each other. Our contributions are:

- A controlled comparison of **four encoders** inside the *same* diffusion policy: ResNet34 with spatial-softmax (the standard CNN baseline) and DINOv2 ViT-S/14 at three state-dropout settings (p=0, 0.2, 1.0), with a fair same-dropout pair (ResNet34-sd02 vs. DINOv2-sd02).
- A **representation-probe suite**, decoder feature inversion, a physical pose probe (linear vs. MLP), and a modality-attribution score, that measures what each encoder retains rather than only end-to-end accuracy.
- A **needle-localization query test** that is, to our reading, the most deployment-relevant finding: DINOv2 patch similarity traces the needle while ResNet34 does not.
- A **state-dropout sweep** that exposes a residual state-shortcut in DINOv2 at p=0 and removes it as p increases, converting a state-coupled policy into a genuinely vision-grounded one.
- **First-hand real-robot observations** (clearly marked as qualitative) that corroborate the offline probes in closed-loop control.

2 Related Works

Diffusion policies. Diffusion policy (Chi et al., 2023) casts visuomotor imitation as conditional action denoising, building on denoising diffusion probabilistic models (Ho et al., 2020) and accelerating inference with the implicit (DDIM) sampler (Song et al., 2021). We adopt this formulation unchanged and vary only the visual encoder and the state-dropout regularizer.

Visual encoders. ResNet (He et al., 2016) with a spatial-softmax keypoint read-out (Levine et al., 2016) is the canonical perception front-end for end-to-end visuomotor learning and for the original diffusion-policy implementation (Chi et al., 2023). Vision transformers (Dosovitskiy et al., 2021) replace convolution with patch attention; self-supervised variants such as DINO (Caron et al., 2021) and DINOv2 (Oquab et al., 2024) learn patch features with strong emergent localization, which motivates testing whether they better expose fine surgical structures than a convolutional baseline.

Probing and feature inversion. Linear classifier probes (Alain and Bengio, 2017) are a standard tool for asking what information a representation makes linearly accessible; we extend the idea to a pose regression probe (linear vs. MLP). Feature inversion through a learned decoder, here using the LeRobot world-model utilities (Lin, 2024), measures how much pixel-level information the encoder retains.

Platform and data. Experiments use the da Vinci Research Kit (Kazanzides et al., 2014) with data recorded and loaded in the LeRobot v3 format (Cadene et al., 2024).

3 Methodology

3.1 Task and dataset

We study a dVRK (Kazanzides et al., 2014) suturing task (needle handling). The dataset is stored in LeRobot v3 (Cadene et al., 2024) and contains 84 episodes totalling 43 890 frames at 10 FPS. Each frame provides a single endoscope RGB view at 576×720 resolution (AV1-encoded). The observation is the image plus a 20-dimensional proprioceptive state (PSM1 and PSM2 position, 6D rotation, and jaw angle); the action is a 20-dimensional *anchored-relative* target predicted as a chunk with prediction horizon H=16.

3.2 Diffusion policy

The controller is a conditional U-Net diffusion policy (Chi et al., 2023). During training we corrupt a ground-truth action chunk $\mathbf{a}_0$ with k steps of Gaussian noise and train a network $\boldsymbol{\epsilon}_\theta$ to predict the added noise, conditioned on the observation embedding $\mathbf{o}$ (Ho et al., 2020):

$$\mathcal{L}(\theta) \;=\; \mathbb{E}_{\mathbf{a}_0,\,k,\,\boldsymbol{\epsilon}}\Big[\,\big\|\,\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\big(\sqrt{\bar{\alpha}_k}\,\mathbf{a}_0 + \sqrt{1-\bar{\alpha}_k}\,\boldsymbol{\epsilon},\; k,\; \mathbf{o}\big)\,\big\|_2^2\,\Big], \tag{1}$$

where $\mathbf{a}_0 \in \mathbb{R}^{20\times H}$ is the clean action chunk, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ is the sampled noise, $k \sim \mathcal{U}\{1,\ldots,T\}$ is the diffusion step, $\bar{\alpha}_k = \prod_{s=1}^{k}\alpha_s$ is the cumulative noise schedule, $\mathbf{o}$ is the conditioning observation embedding (encoded image concatenated with the proprioceptive state), and $\boldsymbol{\epsilon}_\theta$ is the learned denoiser. We use $T{=}100$ DDPM training steps and 16 DDIM (Song et al., 2021) inference steps. The *action validation loss* reported later is Eq. (1) evaluated on the held-out split, and because all reported runs share the anchored-relative action representation, these losses are directly comparable.

3.3 Encoders and the fair-comparison protocol

We compare four encoder configurations inside the identical policy (Table 1): one ResNet34 with spatial-softmax (the standard diffusion-policy encoder (Chi et al., 2023; Levine et al., 2016)) at input resolution 224 also tested with higher resolution but didn't change the layer quality (Figure 1,2,3), and three DINOv2 ViT-S/14 (patch size 14) configurations at resolution 392 that differ only in state dropout. The DINOv2 feature dimension is 1152 (CLS token together with a spatial-softmax over the patch grid), and the ResNet34 spatial-softmax feature dimension is 1024.

State dropout (Sec. 3.5) can change visual grounding on its own, so we mark which configurations are matched. The **fair same-dropout pair** is **ResNet34-sd02 vs. DINOv2-sd02**, both at $p{=}0.2$: same regularizer, different architecture. The **DINOv2 sweep** ($p{=}0$, 0.2, 1.0) holds the architecture fixed and varies only the regularizer, isolating the effect of state dropout. We use the shorthand sd00, sd02, and sd10 for $p{=}0$, 0.2, and 1.0.

Table 1: The four encoder configurations. All run inside one fixed diffusion policy. ResNet34-sd02 vs. DINOv2-sd02 is the fair same-dropout pair; the three DINOv2 rows form the state-dropout sweep. ResNet34 feature dimension 1024 is the spatial-softmax output; DINOv2 is 1152.

Encoder	State-dropout p	Input	Feat. dim	Role
ResNet34 + spatial-softmax	0.2	224	1024	CNN baseline (standard DP encoder)
DINOv2 ViT-S/14	0.0	392	1152	ViT, state kept (sd00)
DINOv2 ViT-S/14	0.2	392	1152	ViT, fair same-dropout pair (sd02)
DINOv2 ViT-S/14	1.0	392	1152	ViT, full vision (sd10)

3.4 Training strategy and implementation details

All four policies share the same training recipe and differ only in the encoder and the state-dropout probability. We optimize the noise-prediction objective of Eq. (1) with AdamW, an MSE loss, and exponential moving averaging of the weights (EMA power 0.75). The action chunk uses observation horizon 4 and prediction horizon 16 (execution horizon 8). Noise is added with a 100-step DDPM schedule at training time and sampled with 16 DDIM steps at inference. The encoder is fine-tuned jointly with the policy at a reduced encoder learning rate. Per-model settings are listed in Table 2.

Table 2: Training hyperparameters. Optimizer AdamW; loss MSE; DDPM 100 train steps, DDIM 16 inference steps; EMA power 0.75; horizons (obs/pred/exec) 4/16/8 for all models. Values taken from the training configuration files.

Encoder	Image	Policy lr	Encoder lr	Batch	Epochs
ResNet34 (sd02)	224	1×10^{-4}	1×10^{-4}	16	500
DINOv2 ViT-S/14 (sd00/sd02/sd10)	392	1.6×10^{-4}	1.6×10^{-5}	64	400

3.5 State dropout

To control how much the policy relies on proprioception, we apply *state dropout*: during training the entire state vector is zeroed with probability p before being concatenated into the conditioning embedding,

$$\tilde{\mathbf{s}} \;=\; (1-m)\,\mathbf{s}, \qquad m \sim \text{Bernoulli}(p), \tag{2}$$

where $\mathbf{s} \in \mathbb{R}^{20}$ is the proprioceptive state, m is a per-sample Bernoulli mask, and $\tilde{\mathbf{s}}$ is the (possibly zeroed) state fed to the policy. Inference is unaffected, because the true state is always provided at test time; state dropout only changes how much the policy is forced to learn from vision during training. We sweep $p \in \{0, 0.2, 1.0\}$: $p{=}0$ keeps the state, $p{=}0.2$ matches the ResNet34 baseline, and $p{=}1$ forces the policy to act from vision alone during training.

3.6 Representation probes

We assess what each encoder retains with four probes.

(i) Decoder feature inversion. Following the LeRobot world-model feature-inversion utilities (Lin, 2024), we train a decoder to reconstruct the input frame from the frozen encoder features and inspect both reconstruction MSE and the reconstructed image; lower MSE and sharper reconstructions mean more pixel-level information is retained.

(ii) Physical pose probe. We regress the ground-truth tool position from frozen features with both a linear map and an MLP and report R^2 (Alain and Bengio, 2017). A high linear R^2 means pose is *linearly* accessible; a gap between linear and MLP indicates that pose is present but encoded non-linearly.

(iii) Modality attribution. We measure how much the policy output depends on vision vs. state by perturbing each input (shuffle baseline) and recording the resulting action shift in millimetres. The vision share is

$$R_{\text{vision}} \;=\; \frac{\overline{\Delta_{\text{image}}}}{\overline{\Delta_{\text{image}}} + \overline{\Delta_{\text{state}}}}, \tag{3}$$

where $\overline{\Delta_{\text{image}}}$ is the mean action shift (mm) when the image is perturbed and $\overline{\Delta_{\text{state}}}$ the mean shift (mm) when the state is perturbed. $R_{\text{vision}} \to 1$ indicates a vision-grounded policy and $R_{\text{vision}} \to 0$ a state-driven one; values near 0.5 indicate a mixed policy. We additionally report the absolute magnitudes, because a high ratio with tiny absolute image sensitivity would not by itself be evidence of useful visual grounding (Sec. 4.4).

(iv) Needle-localization query test. We seed a similarity query patch on the suturing needle and visualize the top-k most-similar patches and a similarity-heatmap overlay across 40 frames of the same clip, at matched resolution between encoders. A good surgical encoder should place its similarity mass on the needle.

Probe sample-size note (honest disclosure). The probe sample counts differ across runs: the DINOv2-sd00 and DINOv2-sd02 probes come from a local low-RAM run (physical $n_{\text{train}} \approx 1252$), the DINOv2-sd10 physical probe used the full data ($n_{\text{train}} = 34\,575$), and the ResNet34 physical probe used $n_{\text{train}} = 953$. The R^2 and R_{vision} values are robust to sample size in this regime, but the counts differ and we flag this where it matters.

4 Results

We treat ResNet34 with spatial-softmax (Chi et al., 2023; Levine et al., 2016) as the baseline and DINOv2 as the candidate encoder. Table 3 collects the comparable per-encoder metrics; cells that were not separately logged are marked n/a.

Note on evaluation metrics. Our task is continuous action *regression* (predicting 20-dimensional anchored-relative motion chunks), not classification, so a confusion matrix is not applicable. In its place we report regression-appropriate metrics: the diffusion validation loss, the per-group mean absolute error (MAE) versus epoch (the regression analogue of an accuracy-versus-epoch curve), and the representation-probe scores (R^2 and R_{vision}).

Table 3: Per-encoder results. Action loss is the held-out value of Eq. (1) (lower is better). Physical R^2 is position, mean of PSM1 and PSM2, linear/MLP. R_{vision} (PSM1/PSM2) is the vision share, Eq. (3). DINOv2-sd02 action val_loss was not separately logged (probes only).

Encoder	p	Img	Best val loss (ep)	Phys. R^2 lin / mlp	R_{vision} (PSM1 / PSM2)	Decoder recon.
ResNet34 (baseline)	0.2	224	0.00436 (300)	0.996 / 0.995	0.882 / 0.879	sharp drop, blurry
DINOv2 sd00	0.0	392	**0.00268 (275)**	0.989 / 0.997	0.763 / 0.767	n/a
DINOv2 sd02	0.2	392	n/a	0.990 / 0.997	0.978 / 0.973	n/a
DINOv2 sd10	1.0	392	0.00326 (300)	0.950 / 0.995	0.996 / 0.993	sharp

4.1 Action accuracy

Among the comparable anchored-relative runs, DINOv2 attains the lowest action validation loss: DINOv2-sd00 $0.002\,68$ (epoch 275) $<$ DINOv2-sd10 $0.003\,26$ (epoch 300) $<$ ResNet34 $0.004\,36$ (epoch 300), as summarized in Table 3. The same ordering holds throughout training and in the per-group validation MAE (rot6d, normalized: both DINOv2 variants stay below the ResNet34 baseline at every checkpoint). Both logged DINOv2 variants therefore beat the ResNet34 baseline on raw action accuracy. The DINOv2-sd02 run was used only for probing and its action val_loss was not separately logged, so it is omitted from this comparison.

4.2 Needle localization (centerpiece)

Our key deployment-relevant finding concerns *what* each encoder localizes. Seeding a similarity query on the suturing needle and following the same 40-frame clip at matched resolution, DINOv2's top-k matched patches trace the thin curved needle across frames, while ResNet34's matches scatter onto tissue blocks and tool bodies and do not localize the needle (Fig. 1). In the per-frame similarity-heatmap overlay (Fig. 2) the contrast is sharpest: DINOv2 is *selective*, concentrating high-similarity heat on the needle and suture wire over a cool background, whereas ResNet34 is *diffuse*, spreading similarity across the tissue field rather than isolating the needle.

This is not a ResNet-depth or layer artifact: ResNet34 was probed at both `layer3` and `layer4` and neither localizes the needle, while DINOv2 finds it directly. For a suturing policy, the needle is precisely the object that must be grasped, oriented, and handed off; an encoder that cannot isolate it is a liability at deployment even when its average action loss is acceptable, because recovery and re-grasping depend on visually finding the needle (Sec. 4.5). Side-by-side videos are linked in Sec. 4.6 (V1, V2).

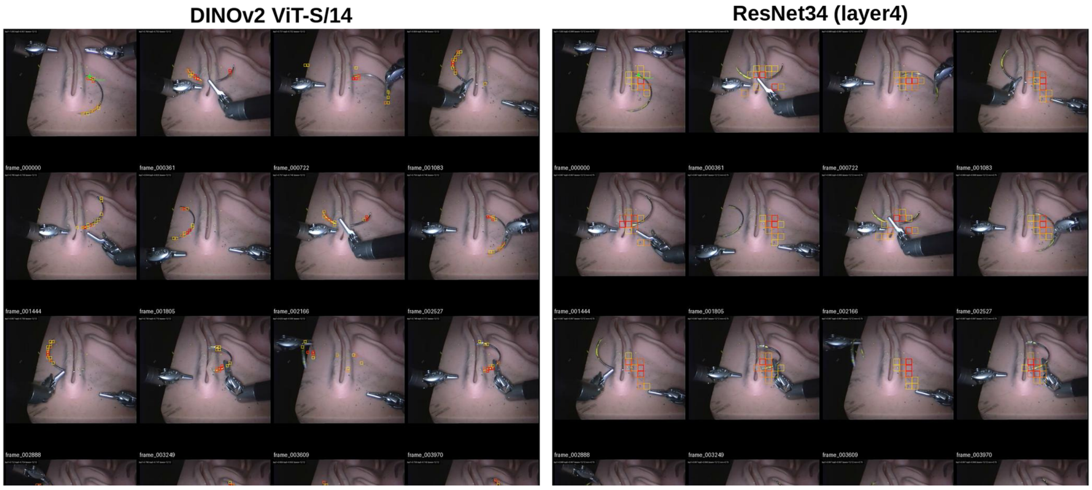

Figure 1: Top-k needle-query patches per encoder (query seeded on the suturing needle; 40 frames). DINOv2 (left) places its top patches on the thin curved needle and suture wire; ResNet34 (right) scatters its top patches over tissue blocks and tool bodies, missing the needle.

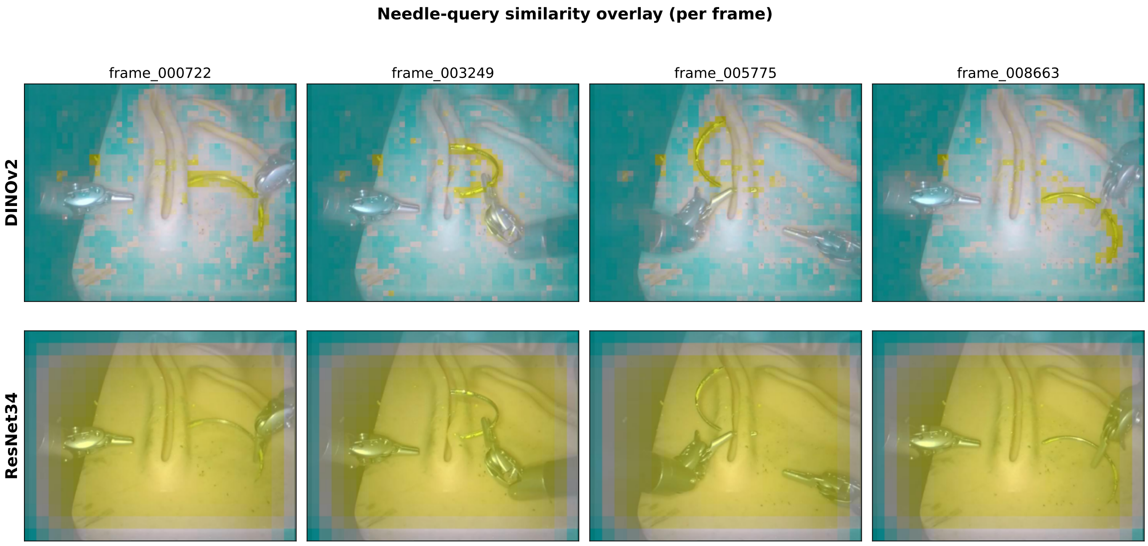

Figure 2: Per-frame needle-similarity overlay. Top row DINOv2, bottom row ResNet34. DINOv2 stays selective with similarity concentrated on the needle and suture across frames; ResNet34 remains non-selective and washes over the tissue.

4.3 Feature retention: what is encoded and how

The encoders also differ in *how* they store task-relevant information.

Decoder feature inversion. The ResNet34 decoder drives reconstruction MSE from 0.0254 (epoch 1) to 0.0022 (epoch 20), so the features clearly retain a lot of pixel structure, yet the reconstruction is visibly *blurry* and loses the needle and other fine detail (Fig. 3, right). The DINOv2 decoder reconstruction is *sharp*, preserving the needle and suture wire (Fig. 3, left). This sharp vs. blurry contrast is the feature-retention counterpart of the needle-localization result: DINOv2 keeps the fine structure that ResNet34 averages away.

Physical pose probe. For physical pose, ResNet34 is *linearly* readable (position R^2 0.996 linear, 0.995 MLP), consistent with its spatial-softmax read-out producing keypoint coordinates. DINOv2 stores pose *non-linearly*: its linear probe is lower and drops further as state dropout increases (sd00 0.989, sd02 0.990, sd10 0.950 linear), but the MLP probe recovers to about 0.997 (sd00/sd02) and 0.995 (sd10), as reported in Table 3. In other words the ViT retains pose at least as well as the CNN but encodes it non-linearly; a linear read-out underestimates what is there. Jaw R^2 is lower for PSM2 across all encoders (for example, ResNet34 linear PSM2 jaw 0.948).

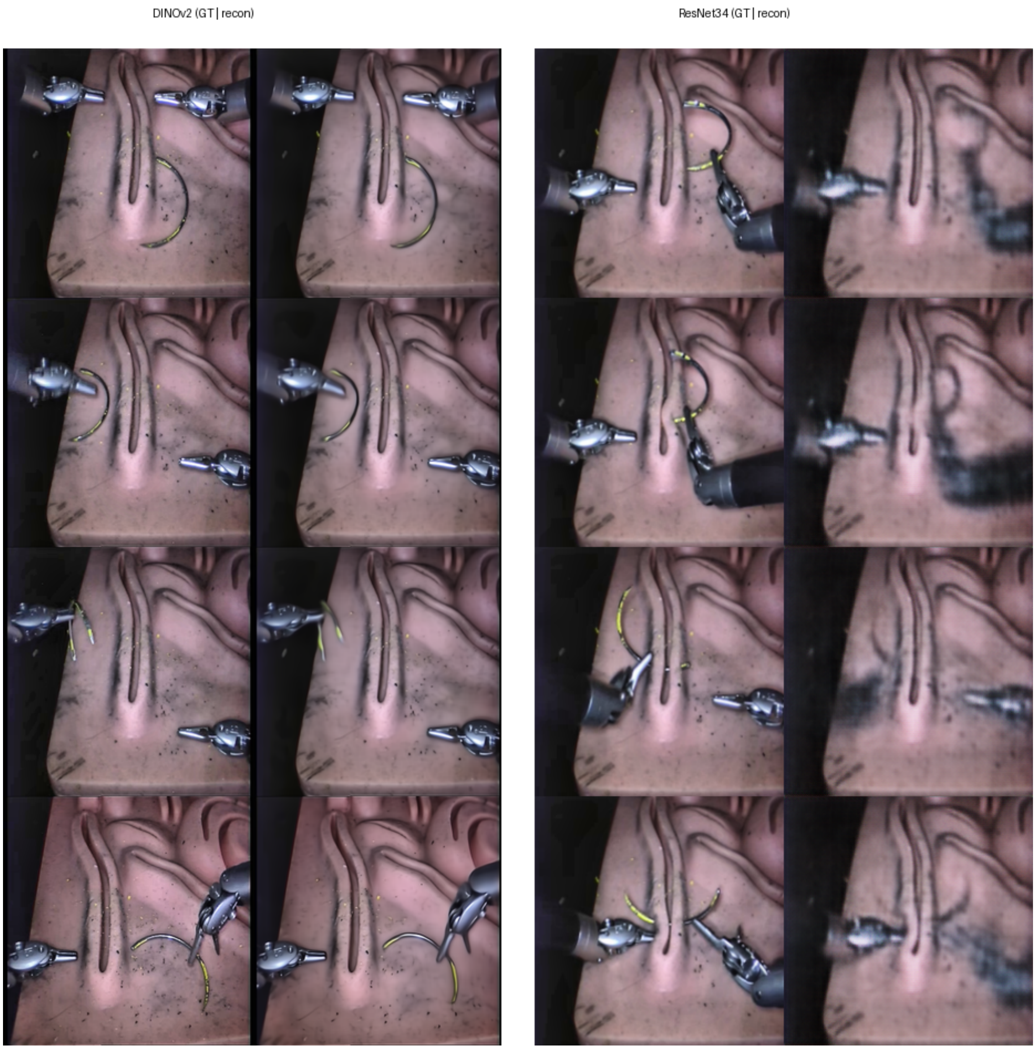

Figure 3: Decoder feature inversion (Lin, 2024). Within each half, the left column is the ground-truth frame and the right column is the reconstruction from the frozen encoder features. DINOv2 (left half) reconstructs the needle and suture wire sharply, whereas ResNet34 (right half) is blurry and loses the needle, even though its reconstruction MSE fell from 0.0254 to 0.0022 over training.

The physical pose-probe scores (R^2, linear and MLP) for all encoders are reported in Table 3, so no separate bar chart is shown here.

4.4 State-dropout sweep and visual grounding

State dropout (Eq. (2)) lets us measure visual grounding directly, and for DINOv2 the sweep is *monotone* and is the headline causal result (Table 4). As p increases, the vision share R_{vision} rises and the action's residual dependence on state shrinks:

- **DINOv2-sd00** (p=0): R_{vision} 0.763 (PSM1, $[0.724, 0.809]$) and 0.767 (PSM2, $[0.705, 0.816]$); image shift 1.519 mm/1.045 mm, state shift 0.472 mm/0.317 mm. This is a *mixed* policy that still uses state, a residual state-shortcut.
- **DINOv2-sd02** (p=0.2): R_{vision} 0.978 (PSM1, $[0.971, 0.983]$) and 0.973 (PSM2, $[0.963, 0.981]$); image shift 1.799 mm/1.198 mm, state shift 0.041 mm/0.033 mm. Strongly vision-grounded.

- **DINOv2-sd10** (p=1): R_{vision} 0.996 (PSM1) and 0.993 (PSM2); image shift 1.752 mm/1.371 mm, state shift 0.007 mm/0.010 mm. Essentially pure vision.

Averaged over the two arms, the residual state sensitivity collapses across the sweep, from roughly 0.39 mm (p=0) to 0.04 mm (p=0.2) to 0.008 mm (p=1), while the image sensitivity stays above 1 mm throughout (Table 4). State dropout thus converts the p=0 state-shortcut into genuine visual grounding without sacrificing visual sensitivity. ResNet34-sd02, the fair same-dropout comparison, is by contrast *mixed* (R_{vision} 0.882 PSM1, 0.879 PSM2) with wide confidence intervals (PSM1 $[0.827, 0.921]$, PSM2 $[0.688, 0.906]$); at the same regularizer DINOv2-sd02 is far more vision-grounded.

Absolute-magnitude caveat. A high R_{vision} alone is not sufficient: it must be accompanied by a large *absolute* image sensitivity. Here that condition holds, because every DINOv2 configuration has an image shift above 1 mm (Table 4), so the high ratios reflect real visual responsiveness rather than a vanishing denominator.

Table 4: State-dropout sweep for DINOv2 (ResNet34-sd02 shown for reference). As p increases, the vision share R_{vision} rises monotonically and the action shift under state perturbation collapses, while the image-perturbation shift stays above 1 mm: state dropout converts a state-shortcut into genuine visual grounding. Values are the mean of PSM1 and PSM2 from the modality-attribution probe; higher R_{vision} and lower state shift are better.

Encoder	p	R_{vision}	Image shift (mm)	State shift (mm)
DINOv2 sd00	0.0	0.76	1.28	0.39
DINOv2 sd02	0.2	0.975	1.50	0.037
DINOv2 sd10	1.0	0.995	1.56	0.008
ResNet34 (ref)	0.2	0.88	1.37	0.19

4.5 Real-robot observations (qualitative)

The following are the author's **first-hand qualitative observations** from closed-loop deployment on the dVRK. They are deliberately reported as qualitative impressions, not quantitative measurements, and no number in this report is drawn from them; they are included because they corroborate the offline probes. Videos are linked in Sec. 4.6 (V3, V4, V5).

- **ResNet (state-shortcut).** Because it lacks a needle feature, the ResNet policy falls into a state-shortcut no matter how long it is trained. In recovery steps the grasp points are effectively tied to joint state, so even when it cannot pick the needle up it still follows those hard-coded motions; if the needle is moved to a random location that does appear in the dataset, it fails to find it. ResNet does a poor job on the actual task.

- **DINOv2-sd02 (residual shortcut).** This policy normally works well, but an unintended state-shortcut is still present: lifting or raising the suturing pad slightly induces errors, revealing a residual state dependence consistent with its mixed-but-vision-leaning probes.

- **DINOv2-sd10 (state-free vision).** This failure mode did not occur. Even with a completely different image setup the policy can produce actions purely from the image, without relying on state, consistent with $R_{\text{vision}} \approx 0.99$.

4.6 Supplementary videos

Supplementary videos are provided as clickable links.

- **V1.** Resnet state-shortcut frames. https://www.youtube.com/watch?v=RQqh8OGS29A
- **V2.** Resnet Following state needle. https://www.youtube.com/watch?v=7DfvQMrawWA
- **V3.** DINOv2 state-free, full-vision-based. https://www.youtube.com/watch?v=prwcpp9BP9c

4.7 Code availability

The full training, evaluation, and diagnostic code is not attached to this report. The diffusion-policy pipeline and the dVRK dataset were developed in the context of the author's role as a student worker, and the implementation belongs to that work, so it cannot be released publicly here. In line with the course guidance that attaching code is optional, the report instead documents every method in the text and traces every reported number to the underlying training and evaluation result files. The decoder feature-inversion and latent-visualization tooling used in Sec. 4.3 was not written from scratch: we adapted the open-source *le-worldmodel* repository (Lin, 2024) (`https://github.com/linzhongzi/le-worldmodel`), reusing its feature-decoding and visualization infrastructure and integrating it into our codebase to invert the frozen encoder features of the diffusion policy.

5 Discussion and Conclusions

Confounds. Several confounds temper the comparison and are stated openly. First, DINOv2 runs at input resolution 392 while ResNet34 runs at 224, which partly favors DINOv2 on fine structure such as the needle. Second, the probe sample counts differ (Sec. 3.6): DINOv2-sd00 and DINOv2-sd02 used a low-RAM run ($n_{\text{train}} \approx 1252$), DINOv2-sd10 used the full data ($n_{\text{train}} = 34\,575$), and ResNet34 used $n_{\text{train}} = 953$; the R^2 and R_{vision} values are robust here but the counts are not equal. Third, the DINOv2-sd02 action val_loss was not separately logged, so the action-accuracy comparison uses sd00 and sd10 only.

Takeaway. Under a fixed diffusion policy on real dVRK suturing data, DINOv2 (i) achieves a lower comparable action validation loss than ResNet34 ($0.002\,68$ vs. $0.004\,36$), (ii) *localizes the task-critical needle*, which ResNet34 does not at either probed layer, and (iii) under a state-dropout sweep converts a residual state-shortcut into genuine visual grounding (R_{vision} 0.76 at p=0 to 0.995 at p=1, with image sensitivity staying above $1\,\text{mm}$). ResNet34 gives a more linearly readable pose and a strong reconstruction-MSE drop, but its reconstruction is blurry, it misses the needle, and it stays state-coupled. For surgical encoders, *what is localized and retained* matters at least as much as average action error, and on that axis the self-supervised ViT, especially at high state dropout, is the stronger choice here.

6 Reflection

This project connected several themes of the Tools for AI course. Working with DINOv2 made the value of self-supervised ViT features concrete: their emergent patch-level localization, learned without labels, was the single most useful property for a fine surgical target, and it contrasted sharply with the convolutional baseline. Using DINOv2 as a fixed front-end for a downstream policy was itself a transfer-learning exercise, and the linear-vs-MLP pose probe and the decoder feature inversion were direct applications of representation-probing ideas from the course: they showed that the ViT retains pose non-linearly, something a linear read-out alone would have hidden. The clearest lesson was that one headline metric can hide reliance: a policy with low action loss can quietly lean on proprioception, and only the modality-attribution probe and the state-dropout sweep exposed and then removed that state-shortcut. I came to see state dropout not just as a regularizer but as a diagnostic *tool* for forcing and verifying visual grounding. Future work is to log the DINOv2-sd02 action val_loss, to export the DINOv2 decoder MSE curve, to equalize probe sample sizes and input resolution, and to replace the qualitative needle assessment with a labelled localization metric.

References

Alain, G. and Bengio, Y. (2017). Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR) Workshop Track*. arXiv:1610.01644.

Cadene, R., Alibert, S., Soare, A., Gallouedec, Q., Zouitine, A., and Wolf, T. (2024). LeRobot: State-of-the-art machine learning for real-world robotics in PyTorch. https://github.com/huggingface/lerobot.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660.

Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burchfiel, B., and Song, S. (2023). Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851.

Kazanzides, P., Chen, Z., Deguet, A., Fischer, G. S., Taylor, R. H., and DiMaio, S. P. (2014). An open-source research kit for the da Vinci surgical system. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 6434–6439.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research (JMLR)*, 17(39):1–40.

Lin, Z. (2024). le-worldmodel: World models for LeRobot. https://github.com/linzhongzi/le-worldmodel.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. (2024). Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*. arXiv:2304.07193.

Song, J., Meng, C., and Ermon, S. (2021). Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*.