This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material for Adapting Generalizable Visual Geometry Models for Endoscopic 3D Reconstruction

Yanzhe Liu[1,3], Yicheng Yu[2], Junyang Wu[2,3], and Yun Gu[1,3] (✉)

[1] School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai, China
[2] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[3] Shanghai Key Laboratory of Flexible Medical Robotics, Tongren Hospital, Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
geron762@sjtu.edu.cn

This document provides dataset and evaluation protocols, implementation and perturbation details, and fine-grained ablations. The protocol below follows the configuration used for the reported experiments.

1 Experimental Protocol and Implementation Details

1.1 Datasets, Annotations, and Evaluation Scope

We train the adaptation framework using C3VD, C3VDv2, SimColon, SCARED, and EndoDepth. C3VD, C3VDv2, SimColon, and SCARED provide both depth and camera-pose supervision and are used for the quantitative evaluation in the main paper. EndoDepth is used only for depth-related training because complete camera-pose annotations are unavailable. Table 1 summarizes the data sources and their roles.

The main experiments follow a *multi-source endoscopic adaptation* setting: reported dataset domains are represented during training and evaluation uses held-out sequences. The results therefore measure sequence-level generalization within observed domains, not zero-shot transfer to an entirely unseen dataset. All splits are performed at the sequence level so that frames from the same source sequence do not occur in both training and test sets.

When an annotation type is unavailable, its supervised loss is disabled. With binary validity indicators m_D and m_P, the supervised objective can be written as

$$\mathcal{L}_{\text{sup}} = m_D \lambda_{\text{depth}} \mathcal{L}_{\text{depth}} + m_P \lambda_{\text{pose}} \mathcal{L}_{\text{pose}}. \tag{1}$$

Thus, EndoDepth samples contribute to depth supervision but not to direct pose supervision. They remain eligible for pose-encoding consistency and Gram Anchoring because these teacher-student terms do not require ground-truth camera poses.

Table 1: Dataset statistics, available annotations, and usage. Each dataset is partitioned at the sequence level using a 60%/40% train/test ratio.

Dataset	Frames	Depth	Pose	Split	Usage
C3VD	10,015	Yes	Yes	60%/40%	Depth + pose
C3VDv2	67,886	Yes	Yes	60%/40%	Depth + pose
SimColon	37,833	Yes	Yes	60%/40%	Depth + pose
SCARED	28,094	Yes	Yes	60%/40%	Depth + pose
EndoDepth	92,672	Yes	No	60%/40%	Depth only

1.2 Preprocessing and Evaluation Protocol

We use the released or dataset-specific preprocessed RGB, depth, pose, and calibration files without an additional common undistortion stage. Images are rescaled and cropped to 224×224 pixels, and camera intrinsics are updated by the same geometric transform. Training uses dynamic clips of 2–24 frames. The number of frames is sampled uniformly for each batch, and frame indices are sampled from a single source sequence. Consequently, there is no fixed frame interval or clip stride. All adaptation variants use the same sampling and preprocessing pipeline.

For depth evaluation, we report AbsRel and $\delta_{1.25}$. Predictions and ground truth are compared only at finite, positive ground-truth pixels. Zero, non-finite, and padded pixels are invalid. We apply per-frame median scale alignment and first compute frame-level metrics, which are then averaged within each sequence and across the test sequences. All baselines are evaluated using the same resizing, masking, and scale-alignment protocol.

For pose evaluation, we report the RMSE of absolute pose error (APE). Given a predicted pose $\hat{P}_i$, a reference pose P_i, and the trajectory alignment transform T_{align}, we compute

$$E_i = T_{\text{align}} \hat{P}_i P_i^{-1}, \qquad \text{APE}_{\text{RMSE}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \|\text{trans}(E_i)\|_2^2}. \tag{2}$$

We use Sim(3) trajectory alignment and report the mean sequence-level APE RMSE. Each test sequence is evaluated as one ordered trajectory. Therefore, the reported protocol does not stitch independently predicted overlapping clips. The same alignment and aggregation procedure is applied to all compared methods.

1.3 Training and Teacher-Student Configuration

The network is initialized from the public VGGT-1B checkpoint. Only the camera and depth heads are enabled. The point and tracking heads are not used. The training configuration is summarized in Table 2. The student is initialized directly from the public VGGT-1B model weights, and the teacher is initialized as an identical copy of the student. Only the student receives gradients. After

Table 2: Training and teacher-student configuration.

Item	Setting	Item	Setting
Backbone	VGGT-1B	Input size	224×224
Patch size	14	Optimizer	AdamW
Learning rate	2×10^{-5}	Weight decay	0.05
Training epochs	31	Precision	bfloat16
Gradient clipping	1.0	EMA decay	0.999
Pose target	`pose_enc`	Consistency	MSE, $\lambda = 0.075$
Gram layers	0, 7, 15, 23	Gram weight	0.05
Batch/GPU setup	48 images/GPU; no accumulation	Test branch	Student

each optimization step, the teacher is updated by exponential moving average (EMA),

$$\theta_T \leftarrow \mu\theta_T + (1-\mu)\theta_S, \qquad \mu = 0.999. \tag{3}$$

Pose consistency is imposed on `pose_enc` before camera decoding. Gram Anchoring uses tokens from layers 0, 7, 15, and 23, and all teacher targets are detached. No module is listed in the frozen-module configuration: the DINO encoder, aggregator, camera head, and depth head are jointly fine-tuned. The point and tracking heads are disabled and are not part of optimization.

2 Perturbation Settings and Additional Ablations

2.1 Perturbation Settings

The degradation-style pool used for perturbation-enhanced adaptation is defined as

$$\mathcal{S} = \{\text{DDPM noise}, \text{zoom blur}, \text{defocus blur}, \text{motion blur}, \text{contrast}, \text{fog}\}. \tag{4}$$

These transformations modify image appearance while preserving the associated depth and camera-pose annotations. For each clip, the perturbation type and severity are sampled once and shared across all frames, while stochastic noise realizations may vary between frames. Table 3 summarizes the parameterization and intended appearance variation of each perturbation. Unless otherwise specified, both the perturbation-enhanced model and the teacher-student model use the complete six-style perturbation pool.

2.2 Ablation on Perturbation Composition

We evaluate how the composition of the perturbation pool affects depth and camera-pose estimation during perturbation-enhanced supervised adaptation. For compact presentation, we use $\mathcal{D}$ for DDPM noise, $\mathcal{B}$ for zoom, defocus, and motion blur, $\mathcal{N}$ for conventional shot and impulse noise, and $\mathcal{C}$ for the natural-image corruption pool consisting of snow, frost, fog, motion blur, brightness,

Table 3: Parameterization and motivation of the endoscopy-oriented perturbations.

Perturbation	Parameter range	Intended variation
DDPM noise	$t \in \{0, \ldots, 10\}$ of 50 steps	Stochastic imaging degradation
Zoom blur	$s \in \{0, 1\}$	Rapid axial camera motion
Defocus blur	$s \in \{0, 1\}$	Short working distance and defocus
Motion blur	$s \in \{0, 1\}$	Rapid endoscope motion
Contrast	$s \in \{0, 1\}$	Non-uniform illumination
Fog	$s \in \{0, 1\}$	Fluid or lens contamination

Table 4: Ablation of perturbation compositions during perturbation-enhanced supervised adaptation for camera pose estimation (APE RMSE ↓).

Configuration	C3VD	C3VDv2	SimColon	SCARED
None	0.4504	0.6731	0.3155	0.8444
$\mathcal{P}$ (proposed)	0.4176	**0.5990**	0.2936	0.8388
$\mathcal{P} \setminus \mathcal{D}$	**0.4094**	0.6180	0.3181	0.7988
$\mathcal{D} \cup \mathcal{N} \cup \mathcal{B}$	0.4258	0.6691	0.3059	0.8106
$\mathcal{D} \cup \mathcal{C}$	0.4238	0.6283	0.2838	0.8184
$\mathcal{N} \cup \mathcal{B}$	0.4351	0.6624	**0.2292**	**0.7745**
$\mathcal{C}$	0.4226	0.6092	0.2707	0.8459
All	0.4223	0.6279	0.3091	0.8070

and contrast. These groups describe the evaluated configurations and are not necessarily mutually exclusive.

The proposed six-style configuration is denoted by

$$\mathcal{P} = \mathcal{D} \cup \mathcal{B} \cup \{\text{contrast}, \text{fog}\}. \tag{5}$$

The remaining variants examine the removal of DDPM noise and the use of conventional noise, natural-image corruptions, or their combinations.

Tables 4 and 5 show that no perturbation composition is uniformly optimal across all datasets and metrics. Compared with standard endoscopic fine-tuning, the proposed six-style pool consistently reduces the pose error on all four datasets. It also reduces AbsRel on C3VD, C3VDv2, and SimColon, although its depth improvement is not consistent on SCARED.

Removing DDPM noise improves pose estimation on C3VD and SCARED but degrades it on C3VDv2 and SimColon, indicating that diffusion-style noise is beneficial only for selected domains. Conventional noise combined with blur achieves the best pose results on SimColon and SCARED and provides strong depth performance on C3VD and C3VDv2. Natural-image corruptions perform well on selected depth metrics but do not yield consistent pose improvements. Moreover, combining all perturbations does not further improve performance, suggesting that excessive perturbation diversity may introduce appearance variations that are less aligned with the target endoscopic domains. We therefore select the proposed six-style pool as a balanced configuration rather than optimizing the augmentation policy for a particular dataset or metric.

Table 5: Ablation of perturbation compositions during perturbation-enhanced supervised adaptation for depth estimation (AbsRel ↓ / $\delta_{1.25}$ ↑).

Configuration	C3VD	C3VDv2	SimColon	SCARED
None	0.1030/0.9304	0.0977/0.9261	0.0319/0.9949	0.0580/0.9690
$\mathcal{P}$ (proposed)	0.1012/0.9294	0.0972/0.9276	0.0304/0.9953	0.0636/0.9682
$\mathcal{P} \setminus \mathcal{D}$	0.1026/0.9261	0.0973/0.9267	0.0301/0.9949	0.0582/0.9739
$\mathcal{D} \cup \mathcal{N} \cup \mathcal{B}$	**0.0963**/0.9349	0.0950/0.9320	0.0308/0.9952	0.0588/0.9717
$\mathcal{D} \cup \mathcal{C}$	0.1012/0.9284	0.0980/0.9267	0.0302/0.9954	0.0587/0.9723
$\mathcal{N} \cup \mathcal{B}$	0.0983/**0.9353**	**0.0949**/**0.9324**	0.0303/0.9954	0.0595/0.9722
$\mathcal{C}$	0.1027/0.9275	0.0985/0.9251	**0.0299**/**0.9955**	**0.0561**/**0.9770**
All	0.1015/0.9289	0.0961/0.9290	0.0301/0.9954	0.0617/0.9645

Table 6: Fine-grained ablation of the teacher-student framework for camera pose estimation (APE RMSE ↓).

Variant	C3VD	C3VDv2	SimColon	SCARED
Default model	0.4198	0.6128	**0.2341**	0.8054
w/ L1 pose consistency	0.4289	0.6482	0.2752	0.8639
w/o pose consistency	0.4322	0.6311	0.2533	0.8055
w/o Gram Anchoring	0.4270	**0.6111**	0.2729	**0.7775**
w/o perturbation augmentation	**0.4012**	0.6750	0.2968	0.8043

2.3 Fine-grained Teacher-Student Ablations

The following experiments operate within the complete teacher-student framework rather than the preceding perturbation-only training stage. The default teacher-student model uses MSE-based pose consistency, Gram Anchoring, and the complete six-style perturbation pool. In the variant without perturbation augmentation, all six perturbation types are disabled, and both the teacher and student receive clean input sequences.

Table 6 isolates the effects of the pose-consistency metric, pose-encoding consistency, Gram Anchoring, and perturbation augmentation. Individual components may benefit different datasets, revealing domain-dependent trade-offs.

Replacing the default MSE consistency loss with L1 increases the pose error across all four datasets. Removing pose consistency also degrades performance on C3VD, C3VDv2, and SimColon, supporting the role of representation-level consistency in stabilizing camera-motion estimation. The effects of Gram Anchoring and perturbation augmentation are more dataset dependent. Removing Gram Anchoring improves C3VDv2 and SCARED but degrades C3VD and SimColon. Removing the perturbation augmentation improves C3VD but substantially increases the errors on C3VDv2 and SimColon.

Table 7 reports sensitivity to the consistency and Gram Anchoring weights. When varying one weight, the other is fixed at its default value: $\lambda_{\text{cons}} = 0.075$ and $\lambda_{\text{gram}} = 0.05$. These values are selected as a balanced configuration across datasets rather than being optimized for a single domain.

Table 7: Loss-weight sensitivity for camera pose estimation (APE RMSE ↓). A dash denotes that the corresponding weight is fixed at its default value.

λ_{cons}	λ_{gram}	C3VD	C3VDv2	SimColon	SCARED
0.01	–	0.4138	0.6624	0.2623	0.8214
0.05	–	0.4144	0.6145	0.3209	**0.7916**
0.075	–	0.4198	**0.6128**	**0.2341**	0.8054
0.10	–	0.4140	0.6257	0.2533	0.8309
0.30	–	**0.4124**	0.6409	0.3366	0.8745
–	0.01	0.4434	0.6313	0.2864	0.8133
–	0.05	0.4198	**0.6128**	**0.2341**	**0.8054**
–	0.10	**0.4099**	0.6572	0.2788	0.8245

The selected weights do not optimize every dataset individually but provide a compromise across heterogeneous endoscopic domains. Pose consistency primarily regularizes camera-motion representations, while its influence on depth is indirect and dataset dependent. Gram Anchoring similarly does not improve every metric uniformly. Overall, the fine-grained results support the complementary roles of perturbation augmentation and representation-level consistency, while also indicating that their effects depend on the target endoscopic domain.

3 Method Scope and Limitations

Our contribution adapts consistency regularization to the internal pose representation and multi-layer token statistics of a joint depth–pose model, together with endoscopy-oriented, geometry-preserving perturbations. It does not claim a new teacher-student paradigm in isolation. The evaluation covers held-out sequences from observed domains, does not separate corruptions by type or severity, and does not explicitly model long-term drift or non-rigid deformation. Although the method is evaluated on multiple public benchmarks, prospective clinical validation remains necessary. Its performance may degrade under extreme occlusion or large inter-frame camera displacement, where reliable visual correspondences become difficult to maintain. The teacher is used only in training and adds no inference overhead to the final single-branch model.