CT2X: AN EMPIRICAL STUDY OF PHYSICS-GROUNDED WORLD MODELS FOR RADIOGRAPHIC NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

CT-grounded virtual views can appear plausible despite incorrect three-dimensional alignment, leaving both their geometric reliability and the source of downstream gains unclear. We introduce CT2X, an empirical framework that makes this information flow explicit: patient-specific CT is a persistent 3D world, a source radiograph updates an explicit rigid pose, a known imaging action analytically propagates that pose, and a task renderer produces the target view. CT2X holds the estimated state fixed to compare direct state use with a virtual-landmark round trip, and matches renderer budgets to compare additional computation with an additional real radiograph. We combine controlled experiments on an 80-case CT cohort with evaluations on four public real-image benchmarks, using learned state correction, intensity-based registration, and learned 2D–3D correspondence under dataset-specific evaluation settings. On 224 view pairs from seven held-out patients, state correction reduces mean future translation error from 47.42 to 7.89 mm; direct use of the estimated pose and recovery from virtual landmarks yield numerically equivalent results across three seeds. For this geometric task, the gain is explained by state correction. In a six-object DeepFluoro comparison, adding a second real view reduces one failed object's mean target registration error from 25.91 to 3.05 mm under a matched maximum budget of 3,000 renderer evaluations; additional single-view optimization leaves that failure unresolved. Gains vary across objects and evaluation settings, and action perturbations cause large target errors even from reference source poses. Evaluations on public datasets use reference-derived actions and assess retrospective geometric reliability. Overall, the results show that virtual navigation depends on accurate state estimation and action fidelity; additional computation cannot always substitute for a new observation.

1 INTRODUCTION

Patient-specific three-dimensional imaging and sparse intraoperative radiographs provide complementary spatial information for image-guided interventions. The 3D image describes anatomy but not its current relationship to the camera; the radiograph captures the current projection but collapses depth. Registering these sources supports geometric overlays and virtual target views, with CT–fluoroscopy fusion investigated in applications such as transcatheter aortic-valve intervention (Markelj et al., 2012; Vernikouskaya et al., 2018). For virtual views to support geometric guidance, however, visual plausibility is insufficient: geometric accuracy must also be established. We study how to assess reliability and attribute measured gains to state estimation, virtual representation, computation, or newly acquired observations.

These distinctions matter because a convincing rendering can coexist with an incorrect 3D state. Multiple virtual views of that state are not independent measurements, and improved performance of a downstream algorithm can result from a better state estimate rather than from its virtual representation. Rendering may still provide a useful task interface, but that benefit requires a comparison that holds the state fixed. Likewise, more optimization of an existing radiograph and acquisition of another real radiograph change different resources. Evaluating sparse-view navigation therefore requires controls that separate changes in state, representation, computation, and observation budget.

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

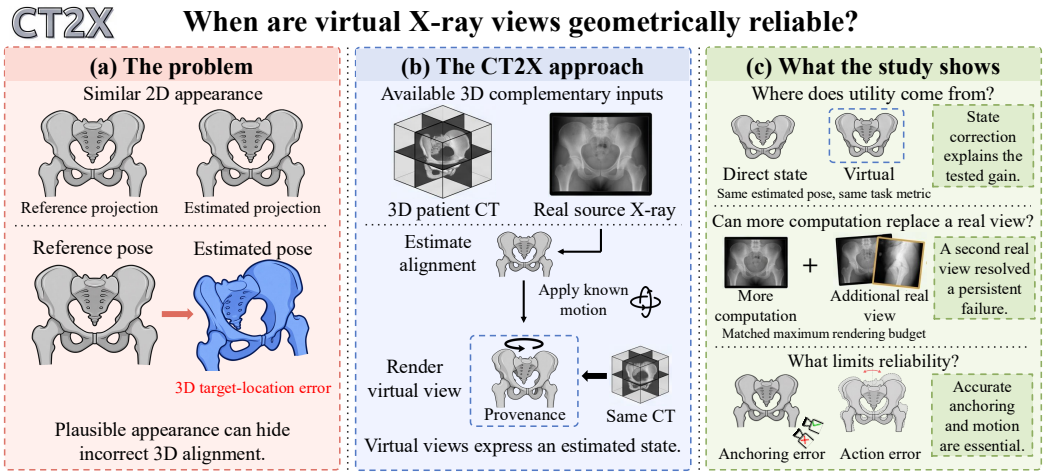


Figure 1: **CT2X overview.** (a) Patient-specific 3D CT and a real source radiograph provide complementary information, yet plausible projections can conceal incorrect 3D alignment. (b) An observation updates an explicit world-to-camera state; a known action left-composes that state, and a task renderer produces a virtual target output. (c) CT2X compares direct and virtual uses of the same state, additional computation and real measurements under matched renderer budgets, and sensitivity to anchoring and action errors. Public actions are reference-derived, and target pixels are held out from anchoring. All anatomical and radiographic panels are conceptual illustrations.

We introduce CT2X as an empirical framework for these comparisons (Figure 1). Following the state–action–observation abstraction of learned world models (Hafner et al., 2019; 2020), the patient-specific 3D image provides a persistent world representation, a source radiograph updates an explicit camera–patient state, an imaging action propagates that state, and a physical renderer produces a target view. Here, the world-model formulation makes each component's information access explicit. X-WIN studies CT-informed representation learning through latent projection prediction (Yang et al., 2026); our setting retains patient-specific structure and analytic geometry to examine the origins and limits of virtual-view utility. Learned state correction, XVR registration (Gopalakrishnan et al., 2026), and RayEmb correspondence with geometric estimation (Shrestha et al., 2025) serve as alternative state estimators within this formulation.

Our study uses an 80-case cardiac CT cohort with synthetic radiographs and four public real-image benchmarks spanning neurovascular and orthopaedic imaging. Public protocols use actions derived from source and target reference poses, while target pixels are held out from anchoring. Fixed protocols expose both successful anchoring and transfer failures. Two comparisons illustrate the value of separating information sources. In the controlled task, direct state use and a virtual-landmark round trip yield the same 7.89 ± 0.32 mm future translation error (mean $\pm$ SD across three model seeds), down from 47.42 mm without correction. The equivalence holds for the tested consumer on all 224 view pairs nested in seven test patients, repeated across seeds. On DeepFluoro, a single-view budget of 3,000 renderer evaluations leaves one failed object at 25.91 mm mean target registration error; a second real constraint under the same maximum renderer budget reduces it to 3.05 mm. Across six objects, this comparison is descriptive and improvements are not uniform.

Our contributions concern evaluation, attribution, and reliability:

1. We specify a world-model evaluation that explicitly tracks real observations, reference geometry, inferred states, and virtual outputs, enabling controlled comparisons of their contributions to task utility.
2. We attribute the controlled consumer's gain to state estimation and show a public-real failure that additional computation leaves unresolved but an additional real constraint repairs under a matched renderer budget.
3. We characterize direction- and object-dependent anchoring, sensitivity to action error, and failures missed by the tested image-similarity and optimization-confidence measures.

The findings motivate separate evaluation of geometric accuracy, action fidelity, and acquisition budget when judging virtual spatial displays. Our evidence concerns retrospective geometric reliability; clinical acquisition savings and clinician benefit remain to be established.

2 Related Work

We organize prior work around the four factors isolated by CT2X: state estimation, additional real observations, registration reliability, and the world-model formulation used to account for information access.

2D–3D registration and state estimation. Rigid 2D–3D registration uses intensity optimization, learned similarity, pose regression, or correspondences (Markelj et al., 2012; Unberath et al., 2021). ProST and DiffDRR provide differentiable projection primitives (Gao et al., 2020; Gopalakrishnan & Golland, 2023). DiffPose and XVR combine patient-specific learning with rendering-based refinement (Gopalakrishnan et al., 2024; 2026), while RayEmb estimates pose from learned landmark–ray correspondences (Shrestha et al., 2025). Chen & Kong (2026) couple spherical similarity learning and an $\mathrm{SO}(4)$-based distance formulation with differentiable Levenberg–Marquardt optimization. DDGS-CT learns a direction-disentangled Gaussian representation for efficient X-ray rendering, improving registration accuracy and runtime in its evaluations (Gao et al., 2024). These advances can improve CT2X's state updater or renderer without making virtual outputs independently acquired observations. Whether our observed single-view failures persist with these tools remains untested; our budget comparison uses a fixed estimator and renderer.

Multi-view registration and additional real observations. Multi-view methods constrain pose through bone-substructure contours (Flepp et al., 2025) or differentiable rendering with cross-view losses (Cui et al., 2025). PolyPose combines sparse real views with anatomical priors through polyrigid transforms for deformable registration (Gopalakrishnan et al., 2025). Its piecewise-rigid model extends beyond CT2X's single global rigid state, illustrating the joint role of observations and structural priors. CT2X instead isolates an additional real constraint from additional computation under a fixed state model and maximum renderer budget; extending this comparison to deformable states remains future work.

Registration reliability and quality assurance. Posterior-based uncertainty quantification studies the nonlinear relationship between uncertainty and registration accuracy (Cho et al., 2025a); learned quality assessment and human–AI verification provide complementary checks (Cho et al., 2025b). CT2X examines failures of image similarity and optimization-confidence measures alongside anchoring and action-induced target error. The resulting audits test whether these surrogates reveal geometric failures; they do not evaluate all uncertainty estimators or human verification protocols.

Physics-grounded world models for predictive sensing. World models learn action-conditioned predictions for control (Hafner et al., 2019; 2020); X-WIN learns chest-radiograph representations through predictive sensing (Yang et al., 2026). CT2X instead retains patient-specific 3D structure, an explicit pose, and analytic rigid transitions to account for information access. We use this formulation to support controlled geometric evaluation, rather than to learn anatomy dynamics or a belief-space acquisition policy (Kaelbling et al., 1998).

3 Physics-Grounded World-Model Formulation

Figure 2 summarizes the state update, analytic propagation, and virtual rendering used in CT2X.

3.1 State, observation, and action

Let M denote a patient-specific 3D intensity volume together with any task-relevant anatomical points or masks. Let $T_t \in \mathrm{SE}(3)$ map the coordinates of M into the camera coordinate system at time or view t, and let K_t contain the calibrated projection geometry. We use this world-to-camera convention throughout, so an action from view t to view $t+1$ left-composes the current state. The

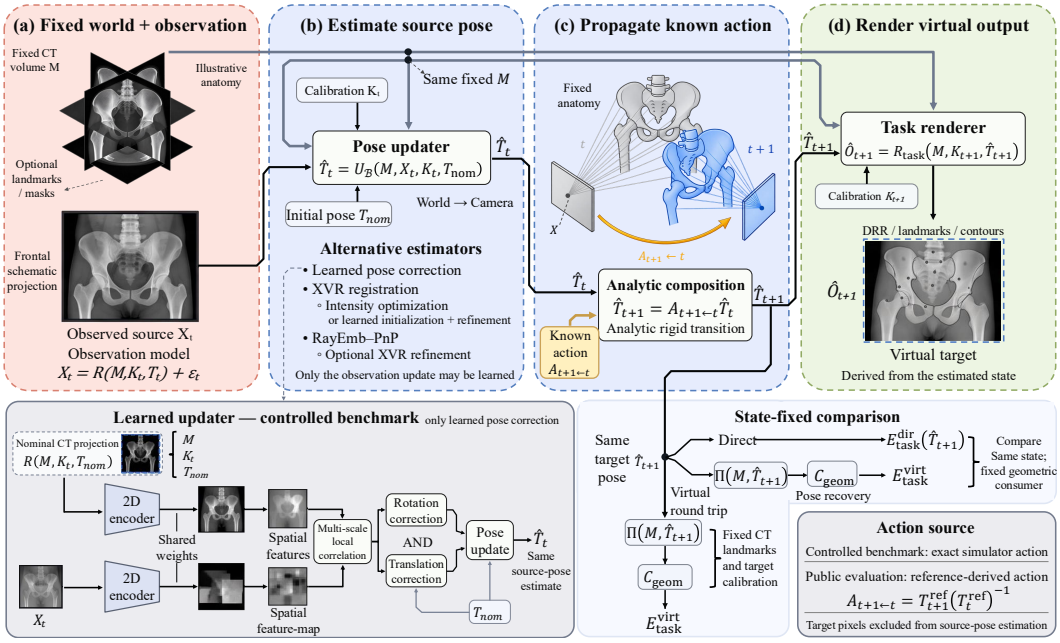


Figure 2: **CT2X framework and evaluation controls.** (a) A fixed patient-specific CT volume and source radiograph support (b) explicit pose estimation, followed by (c) analytic propagation under a known rigid action and (d) virtual rendering from the same CT volume. Lower left: the learned pose corrector used in the controlled benchmark. Lower right: direct pose use and a virtual-landmark round trip share the same estimated target state. Actions are exact in simulation and reference-derived in public evaluations. Anatomy and radiographs are illustrative.

observation model is

$$X_t = R(M, K_t, T_t) + \epsilon_t, \tag{1}$$

where R is an X-ray renderer and ϵ_t collects noise and imaging-model mismatch. M remains fixed in the rigid experiments, while T_t determines its projection. Following learned world models (Hafner et al., 2019; 2020), CT2X separates states, actions, and observations, but fixes M and uses analytic rigid transitions; only the observation update $U_{\mathcal{B}}$ may be learned.

An observation update estimates the state from one acquired source radiograph:

$$\hat{T}_t = U_{\mathcal{B}}(M, X_t, K_t, T_{\text{nom}}). \tag{2}$$

The index $\mathcal{B}$ identifies the inference mechanism. Unlike latent belief-state models that infer a distribution over hidden states from observation–action histories (Igl et al., 2018), CT2X estimates an explicit rigid pose for a fixed patient-specific world. In the controlled benchmark, the learned updater preserves detector-plane feature maps, builds a local multi-scale correlation volume between the observed source image and a nominal CT projection, and regresses a residual rotation and translation. In the public-real experiments, $\mathcal{B}$ is an XVR/DiffDRR intensity optimizer (Gopalakrishnan & Golland, 2023; Gopalakrishnan et al., 2026), the official XVR learned initializer plus refinement, or RayEmb–PnP with optional identical refinement (Shrestha et al., 2025). These are state estimators within the framework, not names for different world dynamics.

Given an action $A_{t+1\leftarrow t} \in \mathrm{SE}(3)$, the next state and output are

$$\hat{T}_{t+1} = A_{t+1\leftarrow t}\hat{T}_t, \tag{3}$$

$$\hat{O}_{t+1} = R_{\text{task}}(M, K_{t+1}, \hat{T}_{t+1}). \tag{4}$$

R_{task} may produce a DRR, projected landmarks, or anatomical contours. It is therefore a task-specific forward operator, not necessarily an intensity renderer. The transition in Equation 3 is analytic; CT2X does not learn anatomy dynamics in the experiments reported here.

3.2 What a virtual observation is—and is not

Conditioned on $(M, X_t, K_t, K_{t+1}, A_{t+1\leftarrow t})$ and a fixed inference procedure, $\hat{O}_{t+1}$ is a deterministic expression of the inferred state. It can be a useful interface for a downstream algorithm or a future human observer, but it is not an independently acquired constraint. To expose this distinction, we compare two paths using exactly the same $\hat{T}_{t+1}$:

$$\text{direct:} \quad \hat{T}_{t+1} \longrightarrow E^{\text{dir}}_{\text{task}}(\hat{T}_{t+1}), \tag{5}$$

$$\text{virtual round trip:} \quad \hat{T}_{t+1} \longrightarrow \Pi(M, \hat{T}_{t+1}) \longrightarrow C_{\text{geom}} \longrightarrow E^{\text{virt}}_{\text{task}}. \tag{6}$$

where Π projects the frozen CT landmarks and C_{geom} is the fixed pose-recovery consumer. $E^{\text{dir}}_{\text{task}}$ and $E^{\text{virt}}_{\text{task}}$ apply the same endpoint metric to the direct and recovered poses. Agreement within the preregistered tolerance establishes equivalence for this state and consumer, attributing the tested gain to state correction; it does not imply equivalence for other consumers or human observers.

3.3 Public-real information boundary

For every public source–target evaluation, the target pixels do not enter Equation 2. However, the action is

$$A_{t+1\leftarrow t} = T^{\text{ref}}_{t+1}(T^{\text{ref}}_{t})^{-1}. \tag{7}$$

We therefore call the protocol *one real source radiograph with a reference-derived known action and held-out target pixels*. It is a controlled test of anchoring and propagation, not evidence that complete patient-coordinate C-arm motion is already available prospectively.

4 Evaluation Design

4.1 Data and protocol roles

Controlled CT2X-Bench. We use 80 locally staged cardiac CT/mask pairs from a TotalSegmentator-derived resource (Wasserthal et al., 2023), split by patient into 64/8/8 training, validation and test cases. The frozen generator produces 32 views and 128 directed view pairs per patient at 256×256 pixels, with a 0.8-mm detector pitch, 1000-mm source–detector distance and 700-mm source–isocenter distance. For the state-perturbation task, the source state is hiddenly perturbed by independently sampled rotations in $\pm 5^\circ$ and translations in ± 10 mm per axis; the exact simulator action is then applied to create the paired target state. The final Gate-V test contains 224 eligible pairs nested in seven test patients. All 32 pairs from the eighth test patient are excluded by the pre-existing requirement of at least four valid landmarks.

Public real images. Ljubljana contains paired 2D/3D cerebral angiography for ten patients (Mitrović et al., 2013). We freeze both frontal-to-lateral and lateral-to-frontal directions with three initializations per patient. DeepFluoro contains calibrated pelvic CT and fluoroscopy from six cadaveric specimens (Grupp et al., 2020); it supports a canonical source protocol, strong estimator baselines, and one fixed four-frame information-budget protocol per object. Femur provides a distinct bone/protocol setting (Flepp et al., 2025). The Yareta hip gold-standard phantom supplies a calibrated markerless boundary test and a fiducial-assisted diagnostic upper bound (D'Isidoro et al., 2021). No pooled metric is computed across these datasets.

Evidence status. We retain the status under which each protocol was frozen. Ljubljana frontal-to-lateral is a confirmatory pass; its reverse direction and the canonical DeepFluoro rule are confirmatory failures under their prespecified criteria. Femur and markerless Yareta are boundary results. Post-hoc observability selection and annotation-assisted runs are explicitly exploratory or upper-bound evidence and appear only in the appendix.

4.2 Compared state estimators

Nominal and reference. The nominal baseline applies the action without correcting T_{nom}. The reference condition starts from the provided or simulated source reference pose and measures propagation error alone.

Learned CT2X state corrector. The controlled learned model encodes the acquired radiograph and a nominal CT projection with a shared 2D encoder, retains a 32×32 spatial feature map, and forms a radius-3 local correlation volume at five image scales $(0.96, 0.98, 1.00, 1.02, 1.04)$. Separate rotation and translation heads estimate a residual correction. Training uses a rotation/translation Smooth-L_1 loss with normalization constants 0.1 radians and 10 mm, plus a landmark-reprojection term weighted by 0.25. We report three frozen checkpoint seeds.

XVR and RayEmb. Our public intensity-based anchor calls the released XVR registration code, which uses DiffDRR; it is not a separate CT2X registration algorithm (Gopalakrishnan & Golland, 2023; Gopalakrishnan et al., 2026). On DeepFluoro we additionally use the official XVR patient-specific learned initializer and test-time refinement. RayEmb uses the official CTPelvic1K-trained checkpoint, 324 templates per object, 3,000 sampled CT points, 600 correspondences, 10 resamplings, and USAC-MAGSAC PnP/RANSAC (Shrestha et al., 2025). We evaluate its PnP estimate alone and after the same frozen XVR refinement. Neither manual source landmarks nor target pixels enter these anchors.

4.3 Representation and information controls

State-fixed representation comparison. We compare direct use of an estimated target pose with projecting its CT landmarks and recovering the pose through the same geometric consumer. Both paths start from the same estimated pose and are evaluated on the same eligible pairs. The preregistered equivalence tolerances are below 0.001 mm for downstream error and pose-translation disagreement, and below 0.001 pixels for projected-landmark disagreement.

Computation and measurement. For the DeepFluoro control, X_1 is the source, X_2 is an additional real constraint, and two X_3 views per object remain held out. The budget-matched single-view optimizer renders X_1 up to 3,000 times. Dual-real jointly optimizes one shared state:

$$\hat{T} = \arg\min_{T} D(X_1, R(M, K_1, T)) + D(X_2, R(M, K_2, A_{2\leftarrow 1}T)), \tag{8}$$

with the same 3,000-render maximum in total; each dual step consumes two renders. The two X_3 images yield repeated fidelity evaluations but not 12 independent specimens.

4.4 Action sensitivity and failure detection

For the controlled action analysis, we propagate the learned pose estimates, the matched physics-based estimates, and the reference source poses with the same perturbation for each pair. We test the clean action and three Gaussian rotation/translation perturbation levels: 2°/5 mm, 5°/10 mm, and 10°/20 mm. The component values are standard deviations; one fixed noise seed is shared across estimators (Appendix C).

The public reliability experiments examine rejection and candidate selection separately. A learned gate accepts or rejects a pose estimate and is evaluated on datasets disjoint from its training data. Coverage is the fraction of accepted estimates; detected failures are rejected estimates whose mTRE meets or exceeds 5 mm. The separate multi-start analysis retains eight candidate poses per anchor. We compare the fixed first candidate with selection by source-image fit. Oracle Recall@K measures the fraction of anchors with at least one candidate below 5 mm among the top K ranked candidates (Appendix D).

4.5 Metrics and statistical units

Public geometric accuracy is the mean target registration error (mTRE) of provided 3D fiducials after propagation. Landmark reprojection error (LRE) is reported separately in detector millimeters for the DeepFluoro information budget and in pixels for the controlled action-noise analysis. Normalized cross-correlation (NCC) measures rendered appearance, but is not treated as a certificate of pose.

For the controlled Gate-V endpoint, E is future-state translation error and the per-pair utility recovery is

$$\mathrm{UR}_i = \frac{E_{\mathrm{nom},i} - E_{\mathrm{method},i}}{E_{\mathrm{nom},i} - E_{\mathrm{ref},i}}. \tag{9}$$

Table 1: Separating representation, computation, and measurement. Panel A uses the controlled Gate-V consumer; errors and recovery are mean ± sample SD across three model seeds. Panel B uses six DeepFluoro objects with two nested target images each. Extra single-real computation and dual-real optimization share a 3,000-render cap; their information budgets differ.

A. Same state, direct versus virtual representation

Condition	Future translation error (mm)	Utility recovery (%)
Nominal state	47.42	0.0
Learned direct state	7.89 ± 0.32	79.50 ± 1.41
Virtual-landmark round trip	7.89 ± 0.32	79.50 ± 1.41
Reference state	0.00	100.0

B. More computation versus an additional real constraint

Condition	mTRE (mm) mean / median	< 2 mm (objects)	< 5 mm (objects)	LRE (detector-mm) mean / median	Runtime (s/object)
Single-real: standard	4.834 / 0.654	5/6	5/6	1.085 / 0.212	3.01
Single-real: extra compute	4.842 / 0.705	5/6	5/6	1.095 / 0.201	63.38
Dual-real: matched cap	1.284 / 0.722	4/6	6/6	0.260 / 0.184	51.89

We average valid per-pair ratios within each checkpoint and report the mean and sample standard deviation across three checkpoints. Consequently, UR is not the ratio of aggregate mean errors. It is a consumer- and reference-specific normalization, not a percentage of information recovered.

The independent unit is always the patient, specimen, or phantom object. Starts, directions, model seeds and held-out target frames are nested repeated evaluations. Where used, confidence intervals resample at the patient/object level. With six DeepFluoro objects, the single/dual analysis remains descriptive even when its tail behavior is striking.

5 Results

5.1 RQ1: State estimation versus virtual representation

We hold the estimated target pose fixed and compare direct pose use with projection to virtual landmarks followed by geometric pose recovery.

State correction reduces mean future translation error from 47.42 mm (nominal state) to 7.89 ± 0.32 mm for both paths (mean ± SD across three checkpoints; Table 1A). All 672 repeated evaluations—224 view pairs from seven patients, evaluated with each checkpoint—meet the equivalence tolerances specified in Section 4.3. The normalized utility recovery in Equation 9 is also identical for both paths. Thus, state correction accounts for the measured gain, while the virtual landmarks preserve the corrected state's accuracy for this geometric task.

5.2 RQ2: How accurate are pose estimates on real radiographs?

Table 2 compares target-view errors and success rates across source directions, datasets, and estimator configurations after propagation using reference-derived actions.

Source direction on Ljubljana. Using three initializations per patient, frontal-to-lateral estimation reaches 1.25 mm mean target mTRE, compared with 6.84 mm in the reverse direction. The fraction of runs below 2 mm falls from 96.7% to 40.0%. The same registration procedure therefore behaves differently when the source and target views are exchanged. Additional rendering metrics are reported in Appendix B.

Failures across datasets. The intensity-based estimator reaches 11.54 mm mean mTRE on Femur, with only one of five objects below 5 mm, and 16.84 mm on the Yareta phantom. Yareta's target-view

Table 2: Frozen public-real target-pose results. Success rates are run-level for Ljubljana and object-level for the other datasets. Groups retain their protocols and evidence status; no pooled ranking is intended.

Direction / estimator	mTRE (mm) mean / median	< 2 mm (%)	< 5 mm (%)	Evidence status
Ljubljana: 10 patients, 30 runs/direction; XVR intensity, 3 starts				
Frontal → lateral	1.246 / 1.349	96.7	100.0	Confirmatory pass
Lateral → frontal	6.837 / 2.566	40.0	93.3	Confirmatory fail
DeepFluoro canonical: 6 objects				
RayEmb–PnP	63.335 / 66.659	0.0	0.0	Frozen strong baseline
RayEmb–PnP + XVR refine	4.794 / 0.672	83.3	83.3	Frozen strong baseline
XVR intensity (2 starts)	4.854 / 0.663	83.3	83.3	Confirmatory fail
Official XVR learned init. + refine	4.901 / 0.746	83.3	83.3	Confirmatory fail
Femur frozen replay: 5 objects				
XVR intensity anchor	11.535 / 10.201	0.0	20.0	Boundary fail
Yareta markerless: 1 object				
XVR intensity anchor	16.843 / 16.843	0.0	0.0	Boundary fail

NCC nevertheless approaches the reference-pose NCC (0.579 versus 0.580), illustrating that similar rendered appearance can accompany an incorrect 3D pose. Supplied fiducials reduce Yareta's error to 1.28 mm in the separate annotation-assisted comparison (Appendix Table 5).

Initialization and refinement on DeepFluoro. RayEmb–PnP alone gives 63.34 mm mean mTRE, with no objects below 5 mm. Adding XVR refinement reduces the mean to 4.79 mm and brings five of six objects below 5 mm. The two-start XVR estimator and official learned XVR initialization followed by refinement reach similar mean errors of 4.85 and 4.90 mm, respectively. All three refined configurations retain the same approximately 26-mm failure on subject04. This persistent failure motivates the next comparison: increasing optimization effort versus adding a second real radiograph.

5.3 RQ3: Computation budget versus measurement budget

We test whether subject04's failure can be resolved by spending more computation on the source image X_1 or by adding a real constraint X_2. The two conditions share a maximum of 3,000 renderer evaluations; accuracy is evaluated at held-out X_3 views (Table 1B).

Joint $X_1 + X_2$ optimization reduces subject04's error from 25.91 to 3.05 mm under the matched 3,000-render cap and brings all six objects below 5 mm. Extra single-view computation changes cohort mean mTRE only from 4.83 to 4.84 mm while runtime rises over twentyfold. The dual-view cohort mean is 1.28 mm (Figure 3); detector-plane LRE and runtime are reported alongside mTRE in Table 1B.

The improvement is concentrated in the failed object. Four of six objects improve, but subject03's error increases from 0.30 to 2.40 mm and the < 2-mm count falls from five to four. The object-level bootstrap mean mTRE change is -3.56 mm, with a 95% interval of $[-11.49, 0.94]$ mm. This six-object comparison identifies a persistent failure resolved by an added measurement under the matched renderer cap, with mixed effects on the remaining objects.

Mean target NCC changes little between extra single-view computation and dual-view optimization (0.457 versus 0.456), despite the reduction in mean mTRE. Appendix Table 6 provides paired case analyses of the directional differences, successful corrections, and persistent failures.

5.4 RQ4: What limits target-view reliability?

Sensitivity to action errors. We apply the same action perturbation to each source-pose estimate and measure landmark mTRE rather than RQ1's translation-only error (Figure 4; Appendix Table 8).

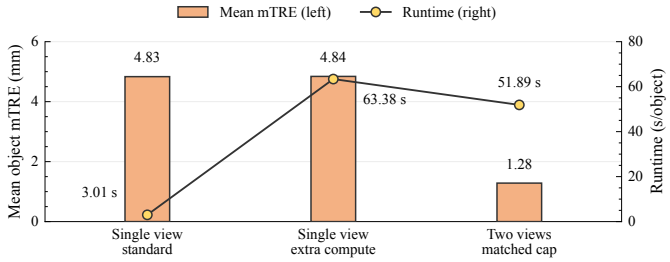


Figure 3: **Computation versus measurement budget on DeepFluoro.** Bars show mean mTRE across six objects (left axis); the line shows runtime per object (right axis). Extra single-view computation and two-view optimization share a maximum of 3,000 renderer evaluations. More computation leaves mean error nearly unchanged, while an additional view reduces it to 1.28 mm.

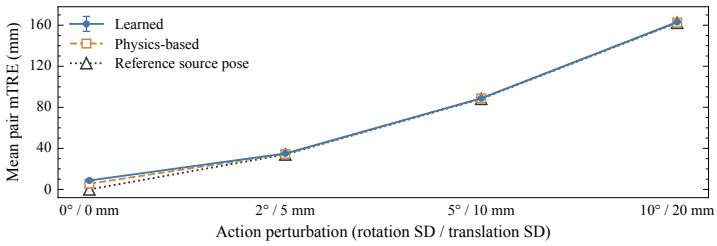


Figure 4: **Same-action error analysis on CT2X-Bench.** Curves show mean mTRE over 224 pairs nested in seven patients. Error bars denote SD over three learned checkpoints. Lines connect the four tested conditions as visual guides; rotation/translation values are Gaussian SDs in the benchmark's fixed coordinate convention.

Without action noise, mean mTRE is 5.56 mm for the physics-based estimator and 8.73 ± 0.28 mm for the learned estimator. At 2°/5 mm noise, the reference source pose alone incurs 34.03 mm target error, close to the 34.57 mm physics-based and 35.02 ± 0.11 mm learned results. The curves remain close at larger perturbations, showing that action error controls the error scale at the tested levels. We use raw mTRE because normalized utility recovery becomes unstable under action noise; Appendix C details this metric choice and the clean-error tails.

Failure detection and candidate selection. Across 32 anchors from 22 patient/object groups, seven fail the 5-mm criterion. The dataset-disjoint acceptance gate retains 59.4% of anchors but detects none of these failures. In the separate eight-candidate analysis, selecting by source-image fit raises success from 68.8% for the fixed first candidate to 78.1%. Searching the top three, top five, or all eight candidates yields the same 78.1% oracle recall: the seven failures contain no successful alternative in the retained pool. These results distinguish two obstacles—failure detection and candidate availability—that remain after optimizing the source image fit (Appendix D).

6 Conclusion

CT2X separates state estimation, virtual representation, computation, and acquired information within an explicit state–observation–action framework. The tested virtual-landmark consumer preserves the estimated state's accuracy, while an additional real observation resolves a failure that extra computation leaves unchanged. Across controlled and public evaluations, target-view reliability depends on accurate anchoring and action fidelity; visual plausibility and optimization agreement alone do not certify physical correctness.

ETHICS STATEMENT

This retrospective computational study uses existing imaging datasets. The controlled cohort comprises 80 cardiac CT/mask pairs from a TotalSegmentator-derived resource. Real-image evaluations use the public Ljubljana, DeepFluoro, Femur, and Yareta datasets under their respective research licenses and attribution requirements. The study involves no prospective intervention, patient contact, or reader study. The case-level audit uses public datasets only. The framework has not been validated for clinical use: visually plausible registrations can be geometrically incorrect, and the reported results do not establish the safety or efficacy of prospective clinical navigation.

AI USE STATEMENT

Generative AI tools were used during the preparation of this manuscript to assist with language polishing and improving the clarity of the writing, to support literature retrieval and discovery of related work, and to help draft portions of the manuscript. All AI-generated content was carefully reviewed, verified, and revised by the authors. The authors take full responsibility for the accuracy, originality, and integrity of the final manuscript.

REFERENCES

Minheng Chen and Youyong Kong. Intraoperative 2d/3d registration via spherical similarity learning and differentiable Levenberg–Marquardt optimization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7239–7249. IEEE, March 2026. doi: 10.1109/WACV61042.2026.00699.

Sue Min Cho, Alexander Do, Robert Grupp, Mehran Armand, Russell Taylor, and Mathias Unberath. Uncertainty Quantification in Image-based 2D/3D Registration and Its Relationship with Accuracy. *International Journal of Computer Assisted Radiology and Surgery*, 20(7):1521–1529, 2025a. doi: 10.1007/s11548-025-03417-x.

Sue Min Cho, Alexander Do, Russell H. Taylor, and Mathias Unberath. Human-AI Collaboration and Explainability for 2D/3D Registration Quality Assurance, 2025b. URL `https://arxiv.org/abs/2507.17597v2`. arXiv preprint.

Yuxin Cui, Rui Song, Yibin Li, Max Q.-H. Meng, and Zhe Min. Robust and Accurate Multi-View 2D/3D Image Registration with Differentiable X-Ray Rendering and Dual Cross-View Constraints. In *IEEE International Conference on Robotics and Automation*, pp. 15451–15457. IEEE, 2025. doi: 10.1109/icra55743.2025.11128375.

Fabio D'Isidoro, Christophe Chênes, Stephen J. Ferguson, and Jérôme Schmid. A new 2d–3d registration gold-standard dataset for the hip joint based on uncertainty modeling. *Medical Physics (Lancaster)*, 48(10):5991–6006, 2021. doi: 10.1002/mp.15124.

Roman Flepp, Leon Nissen, Bastian Sigrist, Arend Nieuwland, Nicola Cavalcanti, Philipp Fürnstahl, Thomas Dreher, and Lilian Calvet. Automatic multi-view x-ray/ct registration using bone substructure contours. *International Journal of Computer Assisted Radiology and Surgery*, 20(7): 1401–1408, 2025. doi: 10.1007/s11548-025-03391-4.

Cong Gao, Xingtong Liu, Wenhao Gu, Benjamin Killeen, Mehran Armand, Russell Taylor, and Mathias Unberath. Generalizing Spatial Transformers to Projective Geometry with Applications to 2D/3D Registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 329–339. Springer International Publishing, 2020. doi: 10.1007/978-3-030-59716-0_32.

Zhongpai Gao, Benjamin Planche, Meng Zheng, Xiao Chen, Terrence Chen, and Ziyan Wu. DDGS-CT: Direction-disentangled gaussian splatting for realistic volume rendering. In *Advances in Neural Information Processing Systems 37*, volume 37, pp. 39281–39302. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-1240.

Vivek Gopalakrishnan and Polina Golland. Fast auto-differentiable digitally reconstructed radiographs for solving inverse problems in intraoperative imaging. In *Lecture Notes in Computer Science*, pp. 1–11. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-23179-7_1.

Vivek Gopalakrishnan, Neel Dey, and Polina Golland. Intraoperative 2d/3d image registration via differentiable x-ray rendering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11662–11672. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01108.

Vivek Gopalakrishnan, Neel Dey, and Polina Golland. PolyPose: Deformable 2d/3d registration via polyrigid transformations. In *Advances in Neural Information Processing Systems 38*, volume 38, pp. 67454–67480. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025. doi: 10.52202/085713-2026.

Vivek Gopalakrishnan, Neel Dey, D. Chlorogiannis, Andrew Abumoussa, Anna M. Larson, Darren B. Orbach, Sarah F. Frisken, and Polina Golland. Rapid patient-specific neural networks for intra-operative x-ray to volume registration. *Nature*, 2026. doi: 10.1038/s41586-026-11045-x. URL https://arxiv.org/abs/2503.16309v2.

Robert B. Grupp, Mathias Unberath, Cong Gao, Rachel A. Hegeman, Ryan J. Murphy, Clayton P. Alexander, Yoshito Otake, Benjamin A. McArthur, Mehran Armand, and Russell H. Taylor. Automatic annotation of hip anatomy in fluoroscopy for robust and efficient 2d/3d registration. *International Journal of Computer Assisted Radiology and Surgery*, 15(5):759–769, 2020. doi: 10.1007/s11548-020-02162-7.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning Latent Dynamics for Planning from Pixels. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to Control: Learning Behaviors by Latent Imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for POMDPs. In Jennifer Dy and Andreas Krause (eds.), *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2117–2126. PMLR, 2018. URL https://proceedings.mlr.press/v80/igl18a.html.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X.

Primoz Markelj, Dejan Tomazevic, B. Likar, and F. Pernus. A review of 3d/2d registration methods for image-guided interventions. *Medical Image Anal.*, 16(3):642–661, 2012. doi: 10.1016/j.media.2010.03.005.

Uroš Mitrović, Žiga Špiclin, Boštjan Likar, and Franjo Pernuš. 3d–2d registration of cerebral angiograms: A method and evaluation on clinical images. *IEEE Transactions on Medical Imaging*, 32(8):1550–1563, 2013. doi: 10.1109/TMI.2013.2259844.

Pragyan Shrestha, Chun Xie, Yuichi Yoshii, and Itaru Kitahara. RayEmb: Arbitrary landmark detection in x-ray images using ray embedding subspace. In *Lecture Notes in Computer Science*, pp. 241–257. Springer Nature Singapore, 2025. doi: 10.1007/978-981-96-0901-7_15.

Mathias Unberath, Cong Gao, Yicheng Hu, Max Judish, Russell H Taylor, Mehran Armand, and Robert Grupp. The Impact of Machine Learning on 2D/3D Registration for Image-Guided Interventions: A Systematic Review and Perspective. *Frontiers in Robotics and AI*, 8, 2021. doi: 10.3389/frobt.2021.716007.

Ina Vernikouskaya, Wolfgang Rottbauer, Julia Seeger, Birgid Gonska, Jochen Wöhrle, and Volker Rasche. Improved registration of 3d ct angiography with x-ray fluoroscopy for image fusion during transcatheter aortic valve implantation. *Journal of Visualized Experiments*, 2018. doi: 10.3791/57858. Issue 136.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. TotalSegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023. doi: 10.1148/ryai.230024.

Zefan Yang, Ge Wang, James Hendler, Mannudeep K. Kalra, and Pingkun Yan. X-WIN: Building chest radiograph world model via predictive sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6920–6930, June 2026.

A Protocol and Artifact Details

Table 3: Evaluation materials and information conditions. Starts, target images, and model seeds are repeated evaluations, not additional independent subjects. CT2X-Bench uses exact simulator actions. Public protocols use reference-derived actions and hold target pixels out from anchoring.

Dataset	Imaging data	Independent units	Evaluation role
CT2X-Bench	80 cardiac CT/mask pairs; synthetic X-rays	64/8/8 patients; 7 evaluable test patients	Controlled mechanisms
Ljubljana	10 3D angiograms; 20 clinical 2D images	10 patients; 3 starts/direction	Directional public-real test
DeepFluoro	6 pelvic volumes; calibrated fluoroscopy	6 objects	Strong baselines; 1/2-view budget
Femur	5 femur volumes; real radiographs	5 objects	Frozen boundary
Yareta	1 calibrated hip phantom	1 object	Markerless boundary / assisted bound

A.1 Information ledger

Table 4 states which information each protocol makes available. In particular, *held-out target* refers to pixels, not to the reference pose used to construct the public action.

Table 4: Information ledger. A check mark means that the quantity is available to the source-state estimator or propagation chain; *eval* means it is used only to form an action or metric.

Protocol	3D model	Source pixels	Target pixels	Source ref. pose	Target ref. pose
CT2X-Bench learned	✓	✓	no	eval	eval
Public single-real	✓	✓	eval only	eval/action	eval/action
DeepFluoro dual-real	✓	X_1, X_2	X_3 eval only	eval/action	eval/action
Reference/action-only	✓	no	no	✓	action

A.2 Coordinate convention

T_t maps world/volume coordinates to the camera at view t. With column vectors, the reference-derived relative action is

$$A_{t+1\leftarrow t} = T^{\mathrm{ref}}_{t+1}(T^{\mathrm{ref}}_t)^{-1}, \qquad \hat{T}_{t+1} = A_{t+1\leftarrow t}\hat{T}_t.$$

The RayEmb legacy DiffDRR convention and the current XVR convention were connected by a deterministic camera/volume transform. A hard check at the reference pose left 0.000130 mm mTRE and a maximum fiducial-coordinate residual of 1.53×10^{-5} mm; neither image pixels nor reference pose values were used to estimate that convention transform.

A.3 Controlled benchmark generation

CT2X-Bench uses a patient-disjoint 64/8/8 split. Each patient has 32 base views and 128 directed pairs. The state-perturbation protocol samples a hidden source rotation uniformly in $\pm 5^\circ$ per axis and translation uniformly in ± 10 mm per axis, renders the perturbed source, and applies the unchanged base action to obtain the target. The reported Gate-V result uses three frozen checkpoints. Model seeds are repeated estimators; pair records remain nested in patients.

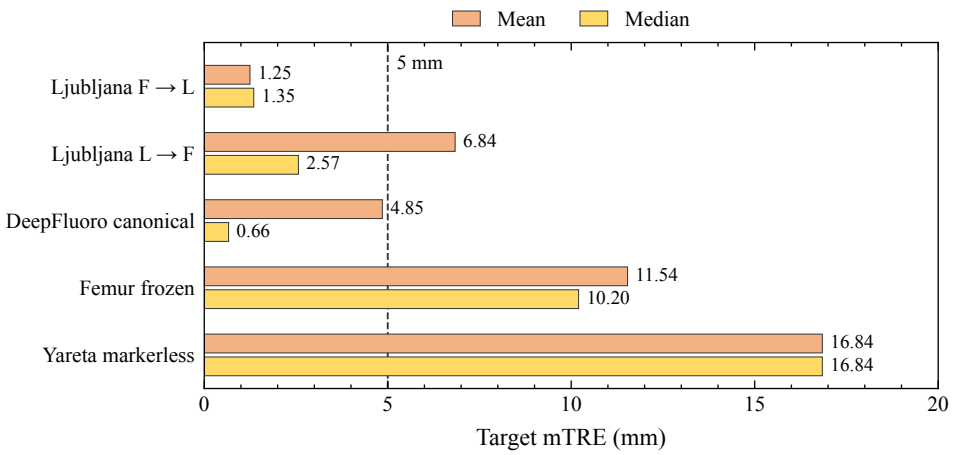


Figure 5: **Geometric accuracy across public datasets.** Orange and yellow bars show mean and median target mTRE. The dashed line marks 5 mm; a mean below this threshold does not imply that every object succeeds, as the DeepFluoro results demonstrate. Rows retain dataset order rather than forming a pooled ranking. All public actions are reference-derived.

B Additional Public-Real Results

Figure 5 contrasts mean and median target errors for the intensity-based estimators across public datasets. Table 2 in the main text gives the corresponding success rates and evidence status.

Rendering accuracy on Ljubljana. For frontal-to-lateral estimation, held-out-view NCC increases from 0.355 to 0.827, corresponding to 96.24% of the oracle-normalized rendering gain in that evaluation.

Table 5: Exploratory, retrospective, and annotation-assisted results excluded from the confirmatory main rows. These values are useful diagnostics but cannot be promoted to unseen confirmation.

Protocol	Mean mTRE (mm)	< 5 mm	Evidence status
DeepFluoro observability-selected v3	0.730	6/6	Exploratory pass
Ljubljana selected-source replay	1.824	10/10	Retrospective external
Yareta supplied-fiducial anchor	1.276	1/1	Annotation-assisted upper bound

DeepFluoro v3 selected a source according to a rule developed after examining the dataset; it is not a replacement for the failed canonical confirmation. The Ljubljana replay applies a subsequently developed selection rule to a previously inspected dataset. Yareta's supplied fiducials inject physical correspondence unavailable to the markerless method.

B.1 Mechanism-defined public-real case audit

We connect cohort summaries to concrete failure mechanisms without choosing aesthetically favorable overlays. This is a post-hoc descriptive audit: the deterministic rule is to report the lowest-index Ljubljana patient in both directions; the DeepFluoro objects that cross a stated 2- or 5-mm status boundary between single-extra and dual-real; and the sole frozen unseen Yareta phantom. DeepFluoro subject03 is also shown under official XVR so that the same object has a concrete successful single-view anchor. This audit is descriptive: rows from the same object and protocol are paired observations, not new sample units.

Table 6: Post-hoc, mechanism-defined public-real case audit (LJ: Ljubljana; DF: DeepFluoro). Cases follow the deterministic rule in Section B.1, which discusses their mechanistic interpretations. The rows are paired within-case diagnostics from different frozen protocols; they are not independent cohort estimates. "Initial to anchored" reports the same rigid state error at the source and reference-action target up to numerical precision.

Case	Frozen contrast	mTRE (mm)	Image / trace diagnostic
LJ-01	F-to-L / L-to-F (8 starts)	0.259 / 3.689	Lateral draw00: 63.933 mm; source-fit MAP picked draw06
DF-03	Official XVR, initial to anchored	65.334 → 0.337	Mean target NCC: 0.397 to 0.446; reference 0.446
DF-04	Official XVR, initial to anchored	85.623 → 26.228	Mean target NCC: 0.498 to 0.509; reference 0.504
DF-04	Single-extra to dual-real (3k-render cap)	25.914 → 3.048	Same render cap; dual objective adds X_2
DF-03	Single-extra to dual-real (3k-render cap)	0.295 → 2.402	Crosses the 2-mm threshold in the adverse direction
Yareta-p21	Frozen two-start markerless policy	16.843	Start disagreement: 0.981 mm; NCC: 0.569 to 0.579; reference 0.580

Ljubljana subject01 shows that direction and initialization can matter within one real clinical patient: frontal anchoring reaches 0.259 mm, whereas the lateral source-fit MAP reaches 3.689 mm; the fixed lateral draw00 is a 63.933-mm failure. DeepFluoro then separates three mechanisms. Subject03 is an accurate official-XVR case (0.337 mm), subject04 remains wrong after the same pipeline even though its mean target NCC exceeds the reference-pose rendering, and a second real constraint moves subject04 from 25.914 to 3.048 mm under the same maximum render count. The paired subject03 deterioration prevents interpreting this rescue as monotonic benefit. Finally, Yareta shows a stable wrong basin: 0.981-mm disagreement and near-oracle target NCC coexist with 16.843-mm target mTRE.

Why private cardiac cases are not shown. A separate readiness audit found only legacy aggregate notes for the former private cardiac-XA experiments; the underlying paired cohort, frozen source/target artifacts, independent device-state reference, and verified ethics/data-use determination were unavailable for audit. We therefore exclude those numbers and images from the evidentiary tables. Table 6 contains public data only; it is a geometric case audit, not a valve-placement or clinician-utility analysis.

B.2 DeepFluoro method and runtime scope

Table 7: Frozen DeepFluoro strong-baseline details. Runtime is shown only with its measured scope; the current two-start result has no comparable runtime.

Method	mTRE (mm) mean / median	Target NCC	Source runtime (s)	Scope
RayEmb–PnP	63.335 / 66.659	0.38580	6.109	PnP pipeline
RayEmb–PnP + XVR	4.794 / 0.672	0.44772	9.526	PnP + test refine
Official XVR	4.901 / 0.746	0.44764	3.22	test only
Current XVR two-start	4.854 / 0.663	0.44768	—	not comparable

The official XVR time excludes patient-specific synthetic-data generation, training, and preprocessing. We therefore do not report a speedup relative to the controlled learned CT2X forward pass or to CPU/SciPy experiments.

Additional information-budget diagnostics. In the single-extra versus dual-real comparison, mean target NCC is 0.456997 and 0.456227, respectively. The object-clustered bootstrap mean mTRE change is -3.558 mm, with a 95% interval of $[-11.490, 0.942]$ mm. Full geometric errors and runtimes appear in Table 1B.

C Action Perturbation Details

Table 8: Same-action target errors. Learned entries are mean $\pm$ sample SD across three checkpoints; physics and reference are deterministic for the fixed perturbation seed. Noise lists rotation/translation standard deviations; LRE is measured in controlled detector pixels.

	mTRE (mm)			LRE (px)		
Noise	Learned	Physics	Reference	Learned	Physics	Reference
0°/0 mm	8.73 ± 0.28	5.56	0.00002	11.03 ± 0.65	5.28	0.00001
2°/5 mm	35.02 ± 0.11	34.57	34.03	57.26 ± 0.26	56.80	56.73
5°/10 mm	88.77 ± 0.21	88.69	88.26	149.77 ± 0.37	149.71	149.38
10°/20 mm	163.13 ± 0.19	162.53	162.34	283.60 ± 0.40	282.97	282.71

The component perturbations are Gaussian standard deviations applied under the frozen benchmark coordinate convention. Rotation center, multiplication order, and translation axes are fixed by the released evaluator.

Clean-error distribution. Without action noise, the physics-based estimator has 5.56 mm mean and 4.09 mm median mTRE, compared with the learned mean of 8.73 ± 0.28 mm. Its 95th-percentile error is 16.12 mm, while the learned 95th-percentile errors range from 13.24 to 14.28 mm across the three checkpoints. The lower physics-based mean therefore coexists with a larger 95th-percentile error.

Metrics under action noise. The same-noise utility-recovery denominator becomes nonpositive for 37, 65, and 78 of 224 pairs at the three nonzero levels: a noisy nominal state can accidentally outperform a noisy reference state. We therefore use raw mTRE and LRE for action-noise evaluation, retaining these pairs and keeping the analysis separate from utility recovery with a clean reference. Because mTRE is a norm, the gaps between the estimator and reference-state curves are not an additive decomposition of state and action error.

D Reliability Audits

The frozen gate audit includes 32 anchors in 22 patient/object groups: 20 Ljubljana directional anchors, six DeepFluoro objects, five Femur objects, and one Yareta phantom. Accept-all has seven failures at 5 mm. The dataset-disjoint learned gate retains 59.4% coverage and detects zero of those seven failures.

For the separate eight-start candidate audit, the fixed first start succeeds on 68.8% of anchors. Selection by source-image fit (called source-fit MAP in the audit) succeeds on 78.1%. Oracle Recall@3, Recall@5, and recall over all eight candidates are also 78.1%; none of the seven MAP failures can be rescued by re-ranking the retained pool. This is a candidate-support result, not a proof that all belief or uncertainty methods must fail.

E Statistical and Reporting Conventions

- The 672 controlled equivalence evaluations are 224 pairs evaluated by three model checkpoints; they are not 672 independent cases.
- Ljubljana's 30 runs per direction are 10 patients times three initializations. Reported percentages are run-level robustness summaries; inferential resampling, where used, is patient-clustered.

- DeepFluoro's two held-out X_3 images per object support image/LRE evaluation but leave six independent objects.
- Mean, median, tail and threshold success are reported together when possible. Mean target error is not used to override a prespecified object-level success criterion.
- Utility recovery is the average of per-pair ratios, not the ratio of aggregate means. Raw physical errors remain primary whenever denominators approach or cross zero.

F CLAIM BOUNDARIES

The experiments do not establish that patient-specific CT is uniquely essential, that complete six-degree-of-freedom state is recovered, that virtual observations create independent information, that confidence is calibrated, that one radiograph is sufficient for surgery, or that CT2X reduces dose. They also do not test clinician accuracy, speed, workload, or safety. A future reader study should first hold geometry fixed to test representation, then use a frozen automatic system to quantify both benefit and the risk of confident geometric errors.