Conformal Safety Certificates in Learned Representations: What Survives and What Must Be Rebuilt

Eric Minwoo Kim **Jong-Kook Kim**
Korea University

Abstract

Real environments are often too complex to model by hand, and robots therefore learn a world model instead. The world model compresses raw sensor data into a compact summary known as a latent state and predicts how that state evolves as the robot acts. Planning then happens inside this learned representation rather than in the physical world. Safety requirements such as clearance are nonetheless defined in the physical world where conformal prediction constructs guaranteed margins in metres, and no existing work examines what those metres mean inside a representation never trained to respect physical distance. This paper argues that such a margin carries three properties that current systems do not distinguish and that only the first survives a change of representation. The three are whether coverage still holds, whether it retains its original physically bounding property, and whether its shape remains useful for planning. A proof-of-concept study calibrates the same margin in physical space, inside the representation, and by a decoder that reads clearance from the latent state, and then repeats every comparison on frozen DINOv2 features as a second and more generic representation. Every margin kept its coverage while regions built on the DINOv2 features hid larger clearance errors and falsely excluded more of the usable space. Only the decoder-calibrated margin remained a bound in metres at no measurable cost in usable space, which suggests that a physically grounded safety guarantee must be built into the world-model interface rather than assumed to survive the change.

1 Introduction

Robots operating in complex environments increasingly learn their world models from data [Ha and Schmidhuber, 2018, LeCun, 2022]. A world model compresses raw observations into a latent state, predicts its evolution during action, and plans entirely inside this learned representation [Ha and Schmidhuber, 2018, Hafner et al., 2019, 2023, Hansen et al., 2024, Zhou et al., 2025, Assran et al., 2025] (Table 1 in Appendix B). Safety requirements have not moved. A drone must keep a stated clearance from a cave wall whatever its internal encoding, in metres rather than embedding distances. Conformal prediction certifies such requirements against imperfect perception by turning an error score and a modest held-out set into a margin that contains the truth at a chosen probability [Vovk et al., 2005, Angelopoulos and Bates, 2021, Lei et al., 2018]. Robotics applies it in physical space to perception [Mei et al., 2026], dynamic-agent forecasts [Lindemann et al., 2023], language-model planners [Ren et al., 2023], and distance fields [Shin et al., 2026, Kim and Kim, 2026]. Recent systems carry the conformal set into latent spaces by conformalizing latent prediction errors [Nath et al., 2026] or calibrating uncertainty thresholds for latent safety filters [Seo et al., 2025]. Each instantiates one particular interface and none examines which certificate properties the representation map preserves. A shared decomposition of the guarantee's meaning after the change is missing.

Three properties are at stake and behave differently. Statistical validity asks whether coverage still holds, physical semantics asks whether membership means anything in metres, and geometry asks whether the certified shape remains tight enough to plan against. Geometry is not a detail. At matched coverage a shape-conforming keep-out reported 1.10 to 1.35 times the usable free space of its convex alternatives and closed-loop safe-reach of 0.98 against 0.70 and 0.54 at fifty flights per arm [Kim and

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Robot Learning with World Models: Capabilities, Frontiers, and Challenges.

Kim, 2026]. Consequences of this size motivate asking what remains when the geometry is learned. This paper decomposes the change of representation into the three properties and sorts existing systems into three design points, physical-first, latent-first, and grounded coupling. A transport lemma and a non-transfer proposition show that exact-image coverage survives any deterministic encoder while the dilation structure behind geometric dominance does not, and geometry must be rebuilt rather than inherited. A proof-of-concept study instantiates every design point on one shared calibration protocol and repeats every comparison on a second and more generic representation. Every margin kept its coverage while the more generic representation hid larger clearance errors and lost usable space, and only the decoder-calibrated margin remained a bound in metres.

2 What survives a change of representation

Setup. Let $\mathcal{X}$ be a physical state space carrying a metric, $\varphi(x)$ true obstacle clearance in metres, $\widehat{\varphi}(x)$ its perceived estimate, $\mathcal{Z}$ a learned latent space, and $E : \mathcal{X} \to \mathcal{Z}$ a deterministic and generally non-invertible encoder. Split conformal calibration takes nonconformity score $A(x)$ and examples $\{x_i\}_{i=1}^n$ exchangeable alongside the test point $x^\star$, and returns the $\lceil (n+1)(1-\alpha) \rceil$-th smallest score δ, which satisfies $\mathbb{P}[A(x^\star) \leq \delta] \geq 1 - \alpha$ free of distributional assumptions [Lei et al., 2018, Angelopoulos and Bates, 2021]. Calibration in $\mathcal{X}$ gives a set $C_\mathcal{X} \subseteq \mathcal{X}$ at level $1 - \alpha$, for instance a keep-out $C_\mathcal{X} = \{x : \widehat{\varphi}(x) \leq r_{\text{safe}} + \delta\}$ whose margin is the quantile of the dangerous error $\max(0, \widehat{\varphi} - \varphi)$ [Kim and Kim, 2026]. Such a certificate has no attachment to planners consuming only latent quantities. Everything following turns on one question. In which space and by which score should A be defined for a planner living in $\mathcal{Z}$?

2.1 Validity transports, semantics and geometry break

Coverage crosses the encoder automatically. For any deterministic E, the exact image $E(C_\mathcal{X}) = \{E(x) : x \in C_\mathcal{X}\}$ inherits the $1 - \alpha$ guarantee, because $x^\star \in C_\mathcal{X}$ implies $E(x^\star) \in E(C_\mathcal{X})$ (Lemma 1, Appendix A). Distortion does not destroy coverage. The two failures lie elsewhere.

Failure (a) — semantic collapse. Nothing prevents E from mapping a safe and an unsafe state to nearby latents, because E was trained for reconstruction or prediction rather than for separating φ level sets. The image $E(C_\mathcal{X})$ then keeps its $1 - \alpha$ coverage while membership stops implying anything about clearance. Coverage is a property of the set's probability mass, safety of its membership function, and the encoder preserves the first while destroying the second.

Failure (b) — unrepresentability. The exact image $E(C_\mathcal{X})$ of a non-convex region generally admits no analytic description, while latent planners impose constraints by representable families such as balls and ellipsoids. Replacing $E(C_\mathcal{X})$ by an approximation $\widehat{C}_\mathcal{Z}$ leaves the lemma inapplicable and the approximation inherits nothing. Coverage must then be re-established in latent space, the latent-first design point and its semantic cost below.

The geometry theorem does not transfer. SCORE's central geometric argument rests on Minkowski dilation, where the certified set $S(t) = O \oplus B(t)$ dilates an obstacle set by a metric ball and preserves the subset ordering of obstacle representations [Kim and Kim, 2026]. Arbitrary encoders destroy it.

Proposition 1 (Non-transfer of dilation structure)**.** *There exist an obstacle set $O \subset \mathcal{X}$, a margin $t > 0$, and a smooth injective encoder E such that for no latent radius $\rho > 0$ does $E\big(O \oplus B(t)\big) = E(O) \oplus B_\mathcal{Z}(\rho)$ hold. It follows that the family $\{E(O \oplus B(t))\}_{t>0}$ is not a Minkowski dilation family in $\mathcal{Z}$, and equal-margin dominance arguments phrased in terms of latent dilation do not follow from their physical counterparts.*

A one-dimensional two-obstacle counterexample proves it (Appendix A). Coverage may transport while the geometry theorem does not unless E preserves the relevant structure. An isometry or a group-equivariant encoder preserves it [Wang, 2026], and generic learned encoders do neither.

2.2 Three design points

The decomposition sorts interfaces into three families that existing systems populate. Ensemble world models such as PETS give calibrated predictive distributions, a model property rather than a finite-sample set certificate [Chua et al., 2018]. Physical certification also reaches beyond statistics to control barrier functions [Ames et al., 2019] and reachability analysis [Bansal et al., 2017], unified as safety filters [Hsu et al., 2024], which face the same crossing.

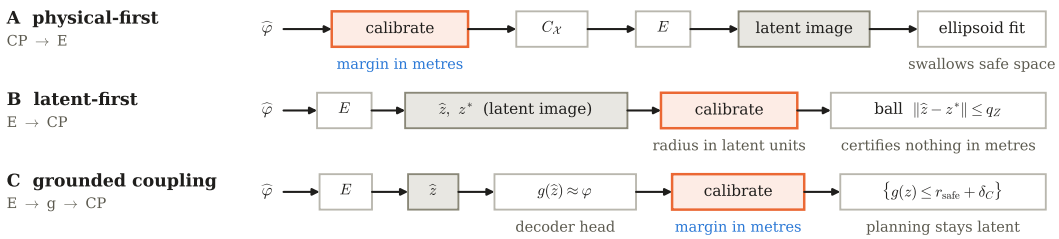


Figure 1: The three design points as pipelines, each beginning from the perceived clearance field $\widehat{\varphi}$. Only Option C recovers a metre-valued quantity after the encoder.

Option A (Figure 1) calibrates in physical space, where recipes are mature [Mei et al., 2026, Lindemann et al., 2023], and pushes the result forward, keeping coverage by Lemma 1 but meeting failure (a), which empties it of meaning, and failure (b), which leaves it unrepresentable. Option B calibrates in the latent space itself. SLS^2 conformalizes latent prediction errors and a learned constraint checker inside robust MPC, the most complete system of this family [Nath et al., 2026], whereas UNISafe calibrates an epistemic-uncertainty threshold and gates out-of-distribution failures by reachability in an augmented latent state [Seo et al., 2025]. Both give immediate validity and a planner-compatible set, but the certified statement carries no physical units and its shape is not intrinsic, because an information-equivalent reparameterization $T : \mathcal{Z} \to \mathcal{Z}'$ warps a Euclidean ball arbitrarily. Option C calibrates the dangerous error $A_{\mathcal{Z}} = \max\big(0, g(\hat{z}) - \varphi(x)\big)$ by a safety head $g : \mathcal{Z} \to \mathbb{R}$ approximating clearance, making δ_C a quantile of clearance overestimates in metres. In world-model vocabulary g is a *decoder head*, decoding the single physical scalar the certificate needs rather than reconstructing pixels, and the keep-out is a sublevel set of latent space that a planner consumes as a differentiable constraint. This mirrors defining the score on a distance field [Kim and Kim, 2026, Shin et al., 2026], where g plays the perceived field's role. Validity and physical units arrive together, but no dominance theorem backs the shape (Proposition 1), closed-loop questions stay open (Section 4), and tightness is empirical (Section 3).

3 Proof of concept on real calibration data

The three options are instantiated on the real calibration data of Kim and Kim [2026], 160 scenes from a DARPA SubT cave environment pairing an accumulated-LiDAR perceived clearance field $\widehat{\varphi}$ and a survey-grade ground-truth field φ, at target coverage 0.90. The encoder E is a small deterministic autoencoder on local patches of $\widehat{\varphi}$ and the head g (Option C only) regresses true clearance, a design kept minimal because the claims concern any deterministic encoder. A pre-registered second encoder, frozen DINOv2 ViT-S/14 features reduced to d=8 and evaluated by identical arms, splits, and estimators (Appendix B), turns each property into a function of encoder genericity. Pairs below are written AE / DINOv2, estimates carry 95% scene-level bootstrap CIs from n=50 test scenes, and the physical margin calibrates to $\delta = 0.71$ m on n=399,704 calibration cells. Each arm reports the metric testing its own weakness, namely safe-cell false-inclusion for failure (a) in A, the metre-spread of dangerous clearance error in B, and the usable-free-space ratio against the native shell in C.

Results. Figure 2 reports the matrix, full values in Table 2. Validity is representation-free, as Lemma 1 predicts. Every calibrated or exact certified object realizes 0.89–0.95 coverage for both encoders. Semantics and geometry are not. Moving to generic DINOv2 features leaves validity flat while the ellipsoid approximation of the pushed-forward set swallows the space (safe-cell false-inclusion $0.108 \to 0.965$), the latent ball's certified cells hide larger physical errors (p95 $0.365 \to 0.554$ m, maximum $0.90 \to 2.95$ m, against Arm A's $\delta = 0.71$ m bound), and rank correlation between latent distance and clearance error falls $0.55 \to 0.21$. A probe shows DINOv2 features still *carry* the task signal (clearance ridge $R^2 = 0.76$, keep-out separability AUC $= 0.99$), making the collapse a representability failure rather than information loss. Arm C keeps its margin in metres ($\delta_C = 0.41/0.43$ m against $\delta = 0.71$ m) and — contrary to the pre-registered expectation of a geometric price — matches the native shell's usable free space at matched coverage (ratios 0.989 and 0.996, CIs spanning 1). Geometry does not survive representation *in* the latent space (Arms A and B). On this data it can be reconstructed *from* the latent by a metric-grounded head at no measurable penalty — premise, not solution. Arm C's DINOv2 coverage dips below target, its CI spanning it (Appendix B).

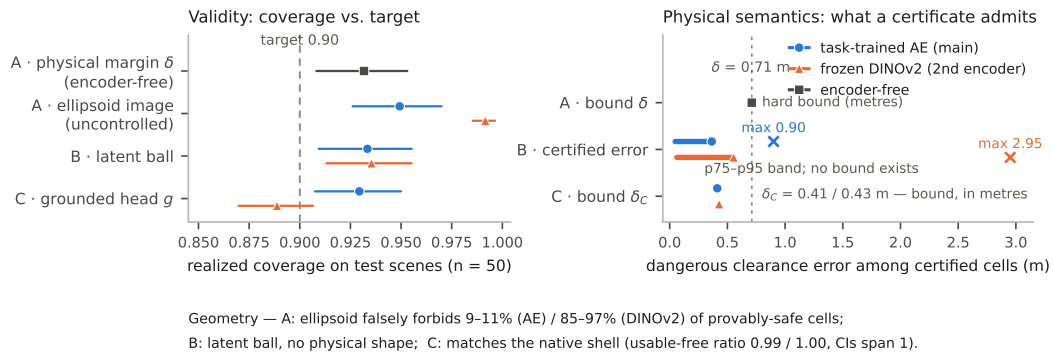


Figure 2: Realized coverage against the 0.90 target (left) and dangerous clearance error in certified cells (right), for both encoders.

4 Desiderata and open problems

This paper's position, restated. *Validity is representation-free but safety is not, and the conformal interface must carry physical units into the representation.* Carrying units is still not carrying geometry. Four criteria judge a satisfying latent conformal safety interface.

1. *Statistical validity.* Nominal coverage holds for the object the planner uses after approximation.
2. *Safety grounding.* Set membership implies a statement in physical units.
3. *Geometric fidelity.* The certified set stays tight enough in the representation where planning happens not to strangle the planner. Proposition 1 blocks the transfer of this property.
4. *Control compatibility.* The planner can efficiently optimize subject to the set, which demands a constraint expressible to CEM, MPPI, or MPC.

Existing systems each satisfy a strict subset. Physical-first designs hold (1)–(2) and lose (3)–(4) at the encoder, latent-first designs hold (1) and (4) by construction, and a grounded head adds (2) and recovers (3) empirically on this data without a guarantee. No known design claims all four, the missing pieces being (3) *by construction* and (4) demonstrated in the loop.

Open problems. (i) *Conditions on E.* What learnable structure (isometry on safety-relevant directions, equivariance [Wang, 2026], level-set preservation of φ) buys geometric fidelity without hurting the world model's primary task? (ii) *Closed-loop exchangeability.* A planner reacting to the certified set shifts deployment away from the calibration distribution [Tibshirani et al., 2019]. (iii) *Rollout compounding.* Multi-step prediction feeds $g(\widehat{z}_{t+k})$ accumulating error and per-step margins need not compose. (iv) *Approximation-aware calibration.* Calibrate the representable set the planner consumes rather than an exact image it cannot. (v) *Non-statistical certificates.* Whether the decomposition extends to barrier conditions and reachable sets, whose transport failure is model mismatch rather than semantic collapse. A natural temporal completion evaluates g on *predicted* rollout latents, where per-horizon margins $\delta_C(k)$ absorb compounding dynamics error, and repurposes the auxiliary decoder to keep calibration honest, auditing executed states against physical measurements and letting margins adapt as deployment shifts (Appendix C).

Limitations. The offline experiment uses two frozen encoders, one dataset, and a 2D clearance field, demonstrating that the three properties separate rather than how every representation behaves. The claims of Section 2 are dimension- and architecture-agnostic while the measurements are not, and Arm C's geometric parity is an observation rather than a theorem. By design no planner consumes the certified sets in the loop, because closed-loop validity is itself open problem (ii).

Conclusion. This paper separated a conformal safety guarantee into statistical validity, physical semantics, and geometry, then traced each property into learned representations. Validity transports automatically, as the transport lemma shows. Semantics and geometry do not, and the non-transfer proposition locates the break in the loss of dilation structure. A three-arm study confirmed the decomposition on real subterranean data and two encoders. Every margin kept its coverage while the more generic representation lost more meaning and usable space, and only the decoder-calibrated margin remained a bound in metres at no measurable cost. Safety guarantees for latent planners should therefore be built into the physically grounded interfaces the desiderata describe.

References

Aaron D. Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control barrier functions: Theory and applications. In *European Control Conference (ECC)*, 2019.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Somil Bansal, Mo Chen, Sylvia Herbert, and Claire J. Tomlin. Hamilton-jacobi reachability: A brief overview and recent advances. In *IEEE Conference on Decision and Control (CDC)*, 2017.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, 2018.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning (ICML)*, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations (ICLR)*, 2024.

Kai-Chieh Hsu, Haimin Hu, and Jaime Fernández Fisac. The safety filter: A unified view of safety-critical control in autonomous systems. *Annual Review of Control, Robotics, and Autonomous Systems*, 2024. arXiv:2309.05837.

Eric Minwoo Kim and Jong-Kook Kim. SCORE: Shape-conforming regions for flight in enclosed, degraded environments. *arXiv preprint arXiv:2608.15289*, 2026.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview preprint*, 2022.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Lars Lindemann, Matthew Cleaveland, Gihyun Shim, and George J. Pappas. Safe planning in dynamic environments using conformal prediction. *IEEE Robotics and Automation Letters*, 8(8): 5116–5123, 2023.

Zhiting Mei, Anushri Dixit, Meghan Booker, Emily Zhou, Mariko Storey-Matsutani, Allen Z. Ren, Ola Shorinwa, and Anirudha Majumdar. Perceive with confidence: Statistical safety assurances for navigation with learning-based perception. *The International Journal of Robotics Research*, 2026. arXiv:2403.08185.

Devesh Nath, Anutam Srinivasan, Haoran Yin, Ruitong Jiang, Jeffrey Fang, and Glen Chou. Pixels to proofs: Probabilistically-safe latent world model control via parallel conformal robust MPC. *arXiv preprint arXiv:2606.15594*, 2026.

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *Conference on Robot Learning (CoRL)*, 2023.

Junwon Seo, Kensuke Nakamura, and Andrea Bajcsy. Uncertainty-aware latent safety filters for avoiding out-of-distribution failures. *arXiv preprint arXiv:2505.00779*, 2025.

Jaeuk Shin, Yoonseok Ra, and Insoon Yang. From prediction uncertainty to conformalized distance fields for safe motion planning. *arXiv preprint arXiv:2607.00776*, 2026.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, 2019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Hongbo Wang. Conformal orbit-valid trust horizons for equivariant world models. *arXiv preprint arXiv:2606.24946*, 2026.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *International Conference on Machine Learning (ICML)*, 2025.

A Proofs

Lemma 1 (Coverage transports along exact images)**.** *For any deterministic E and any measurable $C_{\mathcal{X}}$ satisfying $\mathbb{P}[x^\star \in C_{\mathcal{X}}] \geq 1-\alpha$, the exact image $E(C_{\mathcal{X}}) = \{E(x) : x \in C_{\mathcal{X}}\}$ satisfies $\mathbb{P}\big[E(x^\star) \in E(C_{\mathcal{X}})\big] \geq 1-\alpha$.*

Proof of Lemma 1. $\{x^\star \in C_{\mathcal{X}}\} \subseteq \{E(x^\star) \in E(C_{\mathcal{X}})\}$, since $x^\star \in C_{\mathcal{X}}$ implies $E(x^\star) \in E(C_{\mathcal{X}})$ by definition of the image. Monotonicity of probability gives the claim. No exchangeability beyond what produced $C_{\mathcal{X}}$'s coverage is used. Note the converse inclusion can fail when E is non-injective: $E(x) \in E(C_{\mathcal{X}})$ for $x \notin C_{\mathcal{X}}$ whenever $E(x) = E(x')$ for some $x' \in C_{\mathcal{X}}$ — this surplus membership is exactly the semantic-collapse failure (a) of Section 2. □

Proof of Proposition 1. Take $\mathcal{X} = \mathbb{R}$, $O = \{0, a\}$ with $a > 2t > 0$, and the smooth injection $E(x) = x^3$. Then $O \oplus B(t) = [-t, t] \cup [a-t, a+t]$ and

$$E\big(O \oplus B(t)\big) = [-t^3,\, t^3] \;\cup\; [(a-t)^3,\, (a+t)^3],$$

while for any $\rho > 0$, $E(O) \oplus B_{\mathcal{Z}}(\rho) = [-\rho, \rho] \cup [a^3 - \rho,\, a^3 + \rho]$. Equality of the first components forces $\rho = t^3$. Equality of the second forces $(a+t)^3 - a^3 = \rho$; but $(a+t)^3 - a^3 = 3a^2t + 3at^2 + t^3 > t^3$ for $a, t > 0$. Hence no ρ works, and (the two components being separated, by $a > 2t$, and images/dilations preserving component separation here) the image family is not a latent dilation family. The dilated set's components stretch by different factors under E — precisely the degree of freedom that Minkowski dilation by a single radius cannot express. An isometric E transfers the structure trivially; equivariant encoders transfer it along group orbits [Wang, 2026]; generic learned encoders do neither. □

B Experimental details

Table 1: Planning-time quantities in representative world-model systems. None carries physical units.

System	Representation	Planning-time quantity consumed
DINO-WM [Zhou et al., 2025]	frozen DINOv2 patch latents	CEM on predicted-vs-goal latent distance
TD-MPC2 [Hansen et al., 2024]	learned task latent	MPPI on latent dynamics and learned value
DreamerV3 [Hafner et al., 2023]	RSSM latent	policy and value trained on imagined latent rollouts
PlaNet [Hafner et al., 2019]	RSSM latent	CEM on latent rollout returns

Data and splits. 160 subterranean scenes; per scene, a ± 5 m bird's-eye window with clearance fields φ (survey-grade map) and $\widehat{\varphi}$ (accumulated LiDAR) computed by nearest-neighbour distance on a 0.10 m grid. Scenes are permuted with the source pipeline's seed-0 convention, then split 60/50/50 (train/calibration/test) *by scene*, preventing cell-level leakage; the split deviates from the source's calibration/test split because the encoder and head require training data. Pipeline validation: our loader re-derives the source pipeline's published calibration quantile and shell coverage to float precision before any new measurement. Score band: $0 < \varphi \leq 1.5$ m and $\widehat{\varphi} \leq 3.0$ m, pooled across split scenes; latent quantities use interior cells (full 7×7 patch support).

Learned components. Encoder: MLP autoencoder $49 \to 64 \to 8 \to 64 \to 49$, GELU, MSE, Adam ($10^{-3}$), batch 512, 20 epochs, fixed seed, trained on train-split $\widehat{\varphi}$ patches only, then frozen. Held-out reconstruction RMSE 0.016 m; latent per-dimension standard deviations 0.34–1.99 (no collapse). Head g: $8 \to 32 \to 1$, GELU, Adam (10^{-3}), 20 epochs, train split only; held-out $|g(\widehat{z}) - \varphi|$ RMSE 0.38 m (AE) / 0.43 m (DINOv2).

Second encoder (pre-registered). Frozen DINOv2 ViT-S/14: fields rendered as $\mathrm{clip}(\cdot, 0, 3)/3$ grayscale, resized to 224^2, ImageNet-normalized; final-layer patch tokens ($16 \times 16 \times 384$) bilinearly interpolated to cells; PCA-8 fit on train cells (explained variance 0.54). The keep/drop gate and analysis forks were frozen before any DINOv2 number existed; the drop clause (implementation failure or feature degeneracy) never triggered. Verification probes on the renderings: clearance ridge R^2 — DINOv2-384 0.763, DINOv2-PCA8 0.638, AE-8 0.790; keep-out separability AUC 0.96–0.99. The features carry the task signal; Arm A's collapse under DINOv2 is a representability failure of the explicit set, not information loss.

Ellipsoid fitting (Arm A). Minimum-volume enclosing ellipsoid of the 298,923 calibration keep-out latents via active-set Khachiyan (tolerance 10^{-6}, violator re-insertion), calibration containment > 0.9998 both encoders; a convex-programming (log-det) cross-check on the active set agrees. The spec's ball fallback was not needed.

Estimators. Proportions are pooled per-cell; all CIs are scene-level bootstrap (B=3000, seed 0), re-pooling within each resample; certified-cell error percentiles are recomputed per resample. Geometry uses the source pipeline's matched-coverage protocol: reach sweep 0.10–1.50 m, per-scene coverage of observed-unsafe cells and usable free area (observed cells outside the keep-out), scene-mean curves interpolated at 0.90; the ratio CI bootstraps the whole curve-to-interpolation pipeline. Honest caveat (pre-registered): cells within a scene are dependent; conformal quantiles are pooled-cell (the source convention), CIs at scene level.

Table 2: Property matrix, measured (pre-registered design, 95% scene-bootstrap CIs, B=3000, encoder pairs AE / DINOv2). Validity holds wherever the certified object is calibrated or exact. Each interface then degrades differently, except that Arm C's metric-grounded reconstruction shows no measurable geometric penalty where the pre-registered prediction was a geometric price.

Arm	Validity (cov. @0.90)	Physical semantics	Geometry
A: physical-first, pushed forward	exact image (encoder-free): 0.932 [0.908, 0.953]; ellipsoid approx. uncontrolled: 0.949 [0.926, 0.970] / 0.992 [0.985, 0.996]	margin $= \delta = 0.71$ m by construction; collapse = safe-cell false-inclusion 0.108 [0.090, 0.128] / 0.965 [0.954, 0.975]	image unrepresentable; the fitted ellipsoid forbids 11% / 97% of provably-safe cells
B: latent-first ball	0.933 [0.909, 0.955] / 0.935 [0.913, 0.955]	radius in latent units ($q_{\mathcal{Z}}$=2.7 / 23.0); clearance error among certified cells: p95 0.365 / 0.554 m, max 0.90 / 2.95 m	ball is coordinate-relative, not intrinsic
C: grounded head g	0.929 [0.908, 0.950] / 0.889 [0.870, 0.907]	$\delta_C = 0.41$ / 0.43 m (vs. δ=0.71 m)	no dominance guarantee; measured free-space ratio vs. native shell 0.989 [0.945, 1.004] / 0.996 [0.865, 1.074]

Deviations from pre-registration (reported, not tuned). (1) Arm B's latent-distance/error rank correlation was *moderate* on the AE (0.553 [0.484, 0.622]) where weak was expected; weak on DINOv2 (0.210 [0.139, 0.275]) as registered — moderate correlation still bounds nothing: the certified-cell error maximum (0.90 m) exceeds Arm A's δ. (2) Arm C's keep-out was expected looser than the native shell; measured indistinguishable on both encoders (ratios 0.989 [0.945, 1.004] and 0.996 [0.865, 1.074]). (3) Arm C's DINOv2 realized coverage 0.889 [0.870, 0.907] dips under target with the CI spanning it, consistent with the dependence caveat. No other deviations; all constants as frozen in the spec.

C Toward a bidirectional interface: physical $\leftrightarrow$ latent $\rightarrow$ planner

The static interface demonstrated by Arm C suggests a natural temporal completion. In a latent world-model planner (frozen encoder, learned latent dynamics, auxiliary decoder), the grounded head g would be evaluated not on encodings of current observations but on predicted rollout latents $\widehat{z}_{t+k}$, with per-horizon margins $\delta_C(k)$ calibrated to absorb the dynamics model's compounding prediction error. The decoder, treated in current systems as an optional interpretability tool, becomes load-bearing for *calibration maintenance*: periodically decoding executed states and comparing them

against physical measurements yields fresh $(\widehat{z}, \varphi)$ pairs, allowing the margins to adapt online as closed-loop deployment shifts the visitation distribution away from the calibration data [Tibshirani et al., 2019]. Planning remains entirely latent — the certificate is a differentiable constraint on $\widehat{z}$ — and physical space is consulted only to keep the certificate honest. Validity, physical units, and geometry would then be *maintained* by the interface rather than assumed of the representation. One further lesson of the experiment supports this form: Arm A represented the certified set *extensionally* — an explicit shape in latent coordinates — and collapsed; Option C represents it *intensionally*, as membership through a learned scalar function, which is the form sampling-based latent planners already consume as a cost or constraint [Zhou et al., 2025, Hansen et al., 2024]. This is a sketched research program, not a claim: none of it is demonstrated here.