EviFlow: Structured-Noise Flow Matching for Multi-Rater Segmentation

Anonymous authors
Paper under double-blind review

Abstract

Multi-rater segmentation requires coherent alternative delineations, not merely variable pixelwise predictions. We introduce EviFlow, a conditional flow-matching sampler driven by spatially coherent base noise, together with a distribution-aware evaluation protocol. We establish that, for binary masks on a fixed finite grid, the generalized energy distance (GED) with ground distance $1 - \mathrm{IoU}$ is exactly the squared maximum mean discrepancy under the Jaccard kernel. Its strict positive definiteness makes this GED a strictly proper population divergence; the variant with ground distance $\sqrt{1 - \mathrm{IoU}}$ is strictly proper as well. We distinguish these population guarantees from U- and V-statistic estimation and complement GED with Hungarian-matched IoU (HM-IoU), dispersion, and simple stochastic references. Replacing independent base noise with normalized blurred Gaussian noise yields coherent whole-object alternatives: on LIDC-IDRI with four raters and nine seeds, energy GED decreases from 0.178±0.014 to 0.092±0.022 and HM-IoU rises from 0.361±0.027 to 0.489±0.029. Baseline tuning changes the comparison substantially: a validation-tuned Probabilistic U-Net is not significantly different on energy GED and is better on Jaccard MMD^2 and HM-IoU, whereas its untuned counterpart is the weakest learned stochastic baseline on LIDC GED. On LNDb, EviFlow achieves lower energy GED than the tuned Probabilistic U-Net; additional benchmarks reveal dataset-dependent trade-offs. These findings show the value of structured generative noise and establish a practical case for reporting distributional fit, sample fidelity, dispersion, and baseline tuning together.

1 Introduction

Medical images can admit several plausible segmentations: radiologists may disagree about the boundary, extent, or presence of a faint structure. A distribution over complete masks is therefore a more expressive target than a single consensus annotation. Matching pixelwise foreground frequencies is not enough, because those marginals do not specify which pixels should vary together. Probabilistic segmentation methods address this joint prediction problem by sampling masks conditioned on the image (Kohl et al., 2018; Monteiro et al., 2020).

Evaluation must distinguish distributional fit, sample-to-rater correspondence, and the amount of variation. The generalized energy distance (GED) is widely used for distributional comparison, while Hungarian-matched IoU (HM-IoU) measures agreement under an assignment between sample and rater masks (Kohl et al., 2019). For binary masks on a finite grid, the ordinary IoU-based GED is a strictly identifying population divergence; this follows by combining the energy–kernel equivalence with strict positive definiteness of the Jaccard kernel (Sejdinovic et al., 2013; Bouchard et al., 2013). This connection gives GED a precise distributional interpretation. Complementary sample-quality diagnostics and simple stochastic references then make the practical meaning of an improvement easier to assess.

EviFlow combines an auxiliary evidential marginal head with a conditional flow decoder. Relative to an i.i.d.-noise ablation, spatially coherent base noise produces coherent alternative masks rather than fragmented boundary variation, improving both GED and HM-IoU. The evaluation also exposes a consequential baseline effect: changing one KL weight moves the Probabilistic U-Net from the weakest to the strongest learned stochastic model on LIDC GED. Structured noise and rigorous baseline tuning therefore address different, complementary parts of the problem.

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.

Our contributions are:

- A kernel characterization of multi-rater GED that establishes its strict population guarantees on finite binary-mask spaces and makes the U/V-statistic correction explicit, supported by a complete MinHash proof.
- A structured-noise flow sampler that improves whole-mask coherence, distributional fit, and sample-to-rater agreement, with ablations separating noise structure from the diversity-hinge term.
- An evaluation study showing how baseline tuning and metric choice change method comparisons, together with cross-dataset evidence and practical recommendations for GED, HM-IoU, dispersion, and trivial-sampler references.

2 Related work

Ambiguous segmentation. Probabilistic U-Nets and their hierarchical variants model latent alternatives (Kohl et al., 2018; Kohl et al., 2019; Baumgartner et al., 2019); SSN models spatially correlated logit noise (Monteiro et al., 2020); and MoSE uses stochastic experts (Gao et al., 2023). Diffusion and flow methods generate masks through iterative transformations (Rahman et al., 2023; Zbinden et al., 2023; De Sousa Ribeiro et al., 2025; Li et al., 2026). CAR links calibrated marginals to a stochastic refinement model (Kassapis et al., 2021); our consistency term is inspired by that principle, but uses a one-step surrogate.

Structured noise. Spatial dependence is not new to probabilistic segmentation (Monteiro et al., 2020). Flow-SSN learns the base distribution of a flow (De Sousa Ribeiro et al., 2025), whereas we keep the base construction fixed and image-independent and change its spatial correlation. Related ideas appear in blurring and soft diffusion, inverse heat dissipation, and functional flow matching (Hoogeboom and Salimans, 2023; Daras et al., 2022; Rissanen et al., 2023; Kerrigan et al., 2024). We build on these ideas to study the effect of fixed base-field correlation on multi-rater mask generation.

Evaluation and uncertainty. The energy–MMD correspondence is established (Sejdinovic et al., 2013; Gretton et al., 2012), as are proper scoring rules and energy statistics (Gneiting and Raftery, 2007; Szekely and Rizzo, 2013; Lyons, 2013). The Jaccard index is positive semidefinite (Gower and Legendre, 1986) and strictly positive definite on a finite set universe (Bouchard et al., 2013); we specialize these results to mask evaluation. ValUES emphasizes task-specific validation and strong uncertainty baselines (Kahl et al., 2024). Evidential learning supplies a Dirichlet parameterization (Sensoy et al., 2018), but its vacuity should not automatically be read as faithful epistemic uncertainty (Jurgens et al., 2024). We use temperature scaling and calibration-related losses (Guo et al., 2017; Karandikar et al., 2021), a SegFormer backbone (Xie et al., 2021), and flow matching (Lipman et al., 2023).

3 Population GED and finite-sample evaluation

Setup. Fix an image and a finite pixel universe Ω. A binary mask is a subset of Ω. Let P be a model distribution and Q a target annotation distribution, with independent draws $S, S' \sim P$ and $Y, Y' \sim Q$. For a ground distance d, write the population quantity as

$$G_d(P, Q) = 2\mathbb{E}[d(S, Y)] - \mathbb{E}[d(S, S')] - \mathbb{E}[d(Y, Y')]. \tag{1}$$

The three terms measure cross-fidelity, sample dispersion, and target dispersion. We use G_d for the population divergence and mark finite-sample estimators with subscripts U or V.

Jaccard-kernel identity. Let $k(A, B) = \mathrm{IoU}(A, B)$, with $k(\varnothing, \varnothing) = 1$ and $k(\varnothing, B) = 0$ for nonempty B. Its feature map ϕ satisfies

$$1 - k(A, B) = \tfrac{1}{2}\|\phi(A) - \phi(B)\|_{\mathcal{H}}^2. \tag{2}$$

Substituting into Eq. 1 cancels the constants:

$$G_{1-\mathrm{IoU}}(P, Q) = \|\mathbb{E}_P\phi - \mathbb{E}_Q\phi\|^2_{\mathcal{H}} = \mathrm{MMD}^2_k(P, Q). \tag{3}$$

The mean here is a kernel mean embedding, not a mean segmentation image.

Proposition 1 (Strict identification on a finite mask space). On the masks of a fixed finite pixel universe, $G_{1-\mathrm{IoU}}$ is nonnegative and vanishes exactly when $P = Q$.

Indeed, the full Jaccard Gram matrix K is strictly positive definite. If p and q are the probability vectors over masks, Eq. 3 equals $(p-q)^\top K(p-q)$, which is positive for $p \neq q$. Finiteness makes the corresponding kernel mean embedding injective. Appendix A gives the complete MinHash proof and the precise scope.

Two ground distances. We report $d_r(A, B) = (1 - \mathrm{IoU}(A, B))^r$ for $r = 1$ and $r = 1/2$, called respectively $\beta = 2$ (Jaccard MMD²) and $\beta = 1$ (energy). The latter is proportional to the Hilbert-space distance between feature vectors and identifies distributions because Hilbert spaces have strong negative type (Lyons, 2013). Taking the square root of the ground distance is not taking the square root of the final reported GED. These two divergences can rank imperfect models differently.

U versus V estimation. For $K \geq 2$ samples and $R \geq 2$ annotations, let C average all KR cross-distances and let W_S^U, W_Y^U average within-set distances over distinct indices, not over distinct mask values. Then

$$\widehat{G}_U = 2C - W_S^U - W_Y^U, \qquad \widehat{G}_V = \widehat{G}_U + \tfrac{W_S^U}{K} + \tfrac{W_Y^U}{R}. \tag{4}$$

The U estimator is unbiased under independent, identically distributed draws from each conditional distribution; it can be negative. The V estimator is the divergence between empirical distributions and is nonnegative here, but has finite-sample self-pair bias as an estimator of the population quantity. Only the V estimator is guaranteed to vanish for identical empirical sets. Neither convention removes uncertainty from a small, possibly non-exchangeable panel of human raters.

Complementary diagnostics. HM-IoU maximizes mean IoU under an assignment between samples and tiled annotations; it measures correspondence but is not a proper distributional score. A one-rater example with disjoint masks occurring with probabilities 0.6 and 0.4 gives expected IoU 0.52 when sampling the true distribution, but 0.6 when always predicting the majority mask. We also report the energy-distance dispersion ratio $\rho = W_S^U / W_Y^U$; $\rho \approx 1$ matches average spread, not the full distribution or per-pixel calibration.

Trivial-sampler references and frontiers. A deterministic prediction perturbed by random dilation/erosion or per-pixel Bernoulli noise provides a useful stochastic reference without training another generator. More generally, let $F(w)$ be the lowest GED attained by the tested perturbations at within-sample dispersion w. The diagnostic $\widehat{G}_U - F(W_S^U)$ measures improvement relative to that frontier at matched spread. It complements the proper score rather than replacing it. Our numerical comparison includes a morphological-jitter operating point; we recommend the full dispersion-conditioned frontier when reporting sampler families or diversity sweeps.

4 EviFlow

Shared features and marginal head. A SegFormer-B2 backbone computes $f = \mathrm{Backbone}(x)$. The auxiliary head produces Dirichlet parameters and a marginal:

$$\alpha = 1 + \mathrm{softplus}(W_\alpha f), \qquad \alpha_0 = \sum_c \alpha_c, \qquad \bar{p}_c = \alpha_c / \alpha_0. \tag{5}$$

We combine this marginal with calibration-related training and validation-fitted temperature scaling. The auxiliary vacuity C/α_0 is evaluated separately through risk–coverage analysis.

Conditional flow matching. Let $c = W_c f$ and encode a rater mask as $z_1 = 4(2\,\mathrm{onehot}(y) - 1)$. For base field m_0 and $t \sim \mathcal{U}(0, 1)$, construct $z_t = (1 - t)m_0 + tz_1$ and minimize

$$\mathcal{L}_{\mathrm{flow}} = \mathbb{E}_{t,\, m_0,\, y} \| v_\theta(z_t, t, c) - (z_1 - m_0) \|_2^2. \tag{6}$$

Training uses individual rater masks rather than replacing them with a consensus. At inference, initialize $z^{(0)} = m_0$ and apply

$$z^{(i+1)} = z^{(i)} + \tfrac{1}{N} v_\theta(z^{(i)}, i/N, c), \qquad i = 0, \ldots, N - 1. \tag{7}$$

The final mask is the channelwise argmax; the main setting is $N = 8$ function evaluations and $K = 16$ sampled masks.

The intervention: correlated base noise. The i.i.d.-noise ablation uses $m_0 = \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, I)$ is independent across the spatial grid. EviFlow instead uses

$$m_0 = \frac{\bar{g}_k * \varepsilon}{\mathrm{std}(\bar{g}_k * \varepsilon)}, \qquad k = 17, \tag{8}$$

with a box-blur kernel, used at both training and sampling. We use "GRF noise" as shorthand for the normalized Gaussian-derived field. The base is fixed and image-independent, while the velocity network is image-conditioned. No extra learned parameters are introduced.

Independent noise yields fragmented boundary variation in EviFlow, whereas correlated noise produces coherent whole-object alternatives (Fig. 1). Spatial correlation lets neighboring regions vary together, aligning the base-field structure with the shape-level ambiguity of the task.

Consistency and total loss. We link the two heads through the one-step proxy $\hat{e} = m_0 + v_\theta(m_0, 0, c)$ and $q = (1 + \hat{e}_{\mathrm{fg}}/4)/2$ in a symmetric Bernoulli KL:

$$\mathcal{L}_{\mathrm{cons}} = \tfrac{1}{2}\{\mathrm{KL}(\bar{p}_{\mathrm{fg}} \| q) + \mathrm{KL}(q \| \bar{p}_{\mathrm{fg}})\}. \tag{9}$$

The finite Bernoulli-KL expression is defined on the probability domain $q \in (0, 1)$. The term is a one-step consistency surrogate; it is distinct from matching the final multi-step sample mean exactly to $\bar{p}$. The total objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{edl}} + \lambda_{\mathrm{dice}}\mathcal{L}_{\mathrm{dice}} + \lambda_{\mathrm{flow}}\mathcal{L}_{\mathrm{flow}} + \lambda_{\mathrm{ece}}\mathcal{L}_{\mathrm{ece}} + \lambda_{\mathrm{cons}}\mathcal{L}_{\mathrm{cons}}. \tag{10}$$

The default EviFlow objective omits the diversity hinge. The i.i.d.-noise ablation retains that hinge; a GRF-plus-hinge configuration assesses its contribution separately. These are experimental configurations of EviFlow, not separately named model versions. Architecture and training settings are in Appendix C.

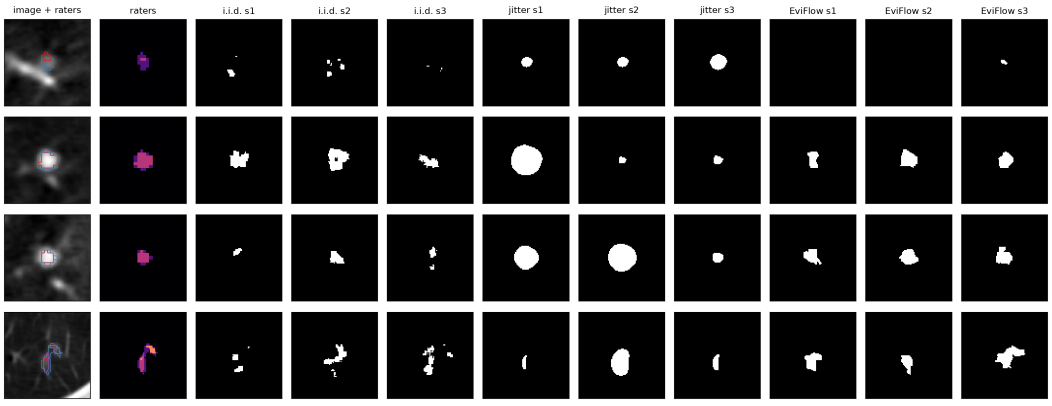

Figure 1. Four high-disagreement LIDC examples. Columns show the image and rater contours, rater mean mask, three i.i.d.-noise ablation samples, three morphological-jitter samples, and three EviFlow samples. The i.i.d.-noise ablation produces fragmented variation, jitter changes the extent of one prediction, and EviFlow produces coherent alternatives. EviFlow's test energy GED is 0.092±0.022, versus 0.152±0.005 for the best reported jitter setting; their HM-IoU values are 0.489±0.029 and 0.494.

5 Experiments

Main protocol. LIDC-IDRI supplies four annotation slots per crop (Armato III et al., 2011). The internal experiment uses patient-disjoint splits, 128×128 crops, a shared SegFormer-B2 backbone, and matched optimization budgets. Table 1 reports means and seed standard deviations, with the number of runs stated for each configuration.

Write the Probabilistic U-Net KL coefficient as b, distinct from the energy-family label β. The validation sweep is $b \in \{0.01, 0.03, 0.1, 0.3, 1, 3, 10\}$ and selects $b = 0.1$ (Appendix C). The $b = 1$ implementation is nearly collapsed, with dispersion 0.15±0.02, versus 0.98±0.03 for the selected model. Reconstruction uses a per-pixel mean, so a numerical KL coefficient cannot be transferred unchanged from a summed-loss implementation. SSN covariance and MC-Dropout rate are also tuned on LIDC validation. This comparison isolates stochastic formulations under a matched backbone; comparisons with original bespoke architectures are identified separately.

Table 1. Internal LIDC comparison at $K = 16$. GED columns use U-statistics; $\beta = 2$ means $d = 1 - \mathrm{IoU}$ and $\beta = 1$ means $d = \sqrt{1 - \mathrm{IoU}}$. HM-IoU uses tiled matching and dispersion uses the energy distance. Values are mean±seed SD; n gives runs. Lower GED and ECE and higher HM-IoU are better; dispersion is compared with 1. The i.i.d.-noise ablation uses independent base noise with the diversity hinge enabled.

Method	n	$\mathrm{GED}_U\ \beta = 2$	$\mathrm{GED}_U\ \beta = 1$	HM-IoU	ρ_U	$\mathrm{ECE}_{\mathrm{band}}$
Deterministic	9	0.525±0.017	0.659±0.026	0.462±0.009	0.000±0.000	0.096±0.003
MC-Dropout	9	0.367±0.017	0.377±0.023	0.495±0.005	0.432±0.025	0.088±0.003
Evidential	9	0.257±0.004	0.238±0.003	0.287±0.012	1.218±0.027	0.125±0.003
Prob. U-Net ($b = 1$, collapsed)	9	0.504±0.016	0.569±0.018	0.462±0.009	0.150±0.019	0.097±0.003
Prob. U-Net ($b = 0.1$, tuned)	9	**0.078±0.009**	**0.081±0.007**	**0.598±0.012**	0.979±0.029	0.119±0.004
SSN (default)	9	0.412±0.021	0.414±0.032	0.480±0.015	0.311±0.102	0.099±0.002
SSN (covariance-tempered)	9	0.260±0.017	0.235±0.016	0.404±0.071	0.959±0.177	0.105±0.006
SSN (rank 50)	3	0.376±0.028	0.365±0.044	0.472±0.019	0.412±0.133	0.101±0.004
i.i.d.-noise ablation	9	0.194±0.015	0.178±0.014	0.361±0.027	1.223±0.058	**0.085±0.006**
EviFlow (ours)	9	0.104±0.022	0.092±0.022	0.489±0.029	1.088±0.056	0.090±0.009
Best trivial sampler (jitter)	–	0.163	0.152±0.005	0.494	0.93	–

Structured noise improves distributional fit and sample fidelity. Relative to the i.i.d.-noise ablation, EviFlow reduces energy GED from 0.178±0.014 to 0.092±0.022 and increases HM-IoU from 0.361±0.027 to 0.489±0.029. Its dispersion moves from 1.22±0.06 to 1.09±0.06. The decomposition shows that the gain comes with lower cross-distances and less excessive spread (Table 3 and Fig. 2). Improvement in both GED and HM-IoU distinguishes better mask alternatives from simply injecting more diversity.

A tuned classical baseline remains competitive. The selected Probabilistic U-Net has energy GED 0.081±0.007, compared with EviFlow's 0.092±0.022; their paired energy-GED difference is not significant after Holm correction ($p = 0.734$). It has lower MMD^2 GED (0.078±0.009 versus 0.104±0.022) and higher HM-IoU (0.598±0.012 versus 0.489±0.029). The MMD^2 and HM-IoU differences favor the tuned baseline (Holm $p = 0.035$ and $p = 0.035$, respectively). The same model at $b = 1$ is the weakest learned stochastic baseline on GED. Thus one diversity-controlling hyperparameter reverses the central ranking in this comparison: tuning is part of a fair evaluation, not an optional detail.

Statistical analysis. Table 2 gives exact paired Wilcoxon tests and Holm correction over nine comparisons within each metric. With nine nonzero differences, the smallest unadjusted two-sided value is $2/2^9$; smaller-seed configurations have coarser attainable values. We distinguish nonsignificant differences from demonstrated equivalence.

Table 2. Paired differences (EviFlow minus comparator), exact two-sided Wilcoxon tests, and within-column Holm correction over nine comparisons. Negative GED and positive HM-IoU differences favor EviFlow. Parentheses show adjusted p and sign agreement k/n. SSN rank 50 has three paired runs.

Baseline	$\Delta D^2_{\mathrm{GED}}\,(\beta=2)$	Holm p (k/n)	$\Delta D^2_{\mathrm{GED}}\,(\beta=1)$	Holm p (k/n)	ΔHM-IoU	Holm p (k/n)
Deterministic	-0.421	0.035 (9/9)	-0.567	0.035 (9/9)	+0.027	0.195 (8/9)
MC-Dropout	-0.263	0.035 (9/9)	-0.285	0.035 (9/9)	-0.006	0.852 (5/9)
Evidential	-0.153	0.035 (9/9)	-0.147	0.035 (9/9)	+0.202	0.035 (9/9)
Prob. U-Net ($b=1$, collapsed)	-0.400	0.035 (9/9)	-0.478	0.035 (9/9)	+0.027	0.219 (7/9)
Prob. U-Net ($b=0.1$, tuned)	+0.026	0.035 (8/9)	+0.010	0.734 (5/9)	-0.109	0.035 (9/9)
SSN (default)	-0.308	0.035 (9/9)	-0.322	0.035 (9/9)	+0.009	0.852 (6/9)
SSN (covariance-tempered)	-0.156	0.035 (9/9)	-0.143	0.035 (9/9)	+0.085	0.035 (9/9)
SSN (rank 50)	-0.284	0.250 (3/3)	-0.285	0.500 (3/3)	+0.027	0.750 (3/3)
i.i.d.-noise ablation	-0.090	0.035 (9/9)	-0.087	0.035 (9/9)	+0.128	0.035 (9/9)

Ablations. GRF noise has the lowest mean GED among the tested base-noise variants, but global and low-resolution noise have five seeds and their pairwise tests against GRF give $p = 0.062$. Adding the hinge back gives 0.104±0.029 versus 0.092±0.022 ($p = 0.250$), with no detected benefit from retaining the hinge. The inference-time scale adjustment M0 worsens energy GED to 0.174±0.042 while reducing dispersion to 0.68±0.12. The soft-Jaccard energy-inspired fine-tune M2 does not improve the observed mean in its single-run exploratory ablation. Structured noise alone therefore provides the strongest result among these tested settings.

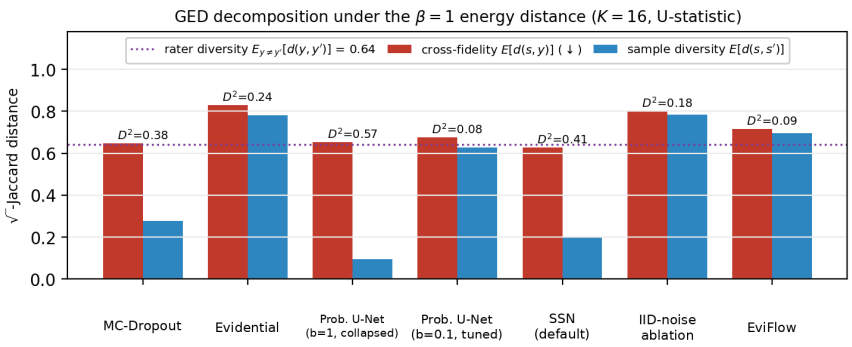

Figure 2. Energy-distance decomposition for the internal LIDC experiment. The red bars show cross-distances, blue bars sample dispersion, and the dotted line rater dispersion. EviFlow reduces both cross-distance and excess spread relative to the i.i.d.-noise ablation, improving GED without increasing diversity.

Table 3. Noise and loss ablations. GED is a U-statistic; cross-fidelity uses $d = \sqrt{1 - \mathrm{IoU}}$. Nine seeds for the i.i.d.-noise ablation, EviFlow (GRF), and GRF+hinge; five for global/low-resolution; three for M0 and one for M2. The ablation p values are unadjusted. Mean values are shown for M0 and M2.

Configuration	$D^2_{\mathrm{GED}}(\beta = 1)$	$D^2_{\mathrm{GED}}(\beta = 2)$	HM-IoU	disp. ratio$_U$	cross-fid. $\mathbb{E}[d(s, y)]$	p vs final
i.i.d. noise, hinge on	0.178±0.014	0.194±0.015	0.361±0.027	1.223±0.058	0.801±0.023	0.004
Global-noise ablation	0.196±0.015	0.220±0.014	0.350±0.060	1.206±0.130	0.805±0.044	0.062
Low-resolution-noise ablation	0.126±0.017	0.149±0.020	0.472±0.031	1.029±0.110	0.713±0.035	0.062
EviFlow (GRF, hinge off)	0.092±0.022	0.104±0.022	0.489±0.029	1.088±0.056	0.715±0.026	–
GRF + diversity hinge	0.104±0.029	0.117±0.030	0.467±0.042	1.124±0.075	0.732±0.037	0.250
+ M0 noise-scale calib.	0.174	0.207	0.513	0.685	0.627	–
+ M2 energy-score FT	0.097	0.107	0.464	1.153	0.738	–

Cross-dataset evidence. On LNDb, EviFlow achieves lower energy GED than the tuned Probabilistic U-Net (0.033±0.046 versus 0.076±0.028, unadjusted paired $p = 0.020$), with no significant HM-IoU difference (Appendix B). EviFlow also has the lowest mean GED in the all-task QUBIQ reference comparison and the official-score Cityscapes comparison. On canonical LIDC, Flow-SSN has the lowest observed GED and the Probabilistic U-Net the highest HM-IoU. Together, these results reveal complementary strengths rather than a single ordering of methods; the dataset-specific protocols and seed counts are given in Appendix E.

The evidential head is auxiliary. The risk–coverage results favor MC-Dropout entropy over evidential vacuity (Table 9), while the per-pixel metrics also vary across methods. The central improvement is in the distribution of complete masks, not a claim of superior epistemic-uncertainty estimation.

Limitations. The mechanism study uses a lightweight velocity network; larger-receptive-field controls and blur-scale sweeps would test its generality. The tuned Probabilistic U-Net remains stronger on LIDC MMD^2 and HM-IoU. Cross-dataset comparisons use different seed counts, and QUBIQ's small, heterogeneous tasks make per-task results important. The binary-mask theory does not establish strict propriety of the official multiclass Cityscapes score or of soft-mask training surrogates.

6 Conclusion

Spatially coherent base noise produces better alternative masks in EviFlow, improving both distributional fit and sample-to-rater agreement. The kernel view explains why IoU-based GED is meaningful for binary-mask distributions, while the empirical comparisons show why it should not stand alone. A single baseline hyperparameter reverses the LIDC GED ranking, and cross-dataset results expose different trade-offs between GED and HM-IoU. We therefore recommend reporting both GED ground distances with explicit U/V conventions, tiled HM-IoU, dispersion, and a trivial-sampler reference or dispersion-conditioned frontier, with every stochastic baseline tuned on validation. These are complementary requirements for assessing meaningful progress in multi-rater segmentation.

Reproducibility statement

Appendix A provides the complete finite-domain proof and explicit estimator and matching conventions. Appendices B–C give dataset counts and training settings, and Appendix E separates the cross-dataset protocols. Tables state their seed counts, estimators, and aggregation rules. The accompanying source package includes machine-readable aggregate results, table-generation code, and independent checks of the finite-mask metric identities.

Ethics statement

The study uses public, de-identified datasets under their respective access and license terms: LIDC-IDRI, LNDb, QUBIQ-2021, and Cityscapes (Cordts et al., 2016). The accompanying package does not redistribute dataset images or annotation collections. Medical segmentation uncertainty is evaluated as a research tool, not as a substitute for clinical judgment or evidence of clinical safety.

Use of Large Language Models

AI assistance supported implementation, literature retrieval, methodological feedback, preparation of tables and figures, aggregate-result checks, mathematical exposition and writing of the finite-domain proof, and drafting and editing text. The authors take responsibility for the final content, including all AI-assisted text, proofs, code, numerical claims, and reproducibility materials.

References

Armato III, Samuel G.; McLennan, Geoffrey; Bidaut, Luc; McNitt-Gray, Michael F.; Meyer, Charles R.; Reeves, Anthony P.; others. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. Medical Physics, 38:915–931. DOI: 10.1118/1.3528204, 2011.

Baumgartner, Christian F.; Tezcan, Kerem C.; Chaitanya, Krishna; Hotker, Andreas M.; Muehlematter, Urs J.; Schawkat, Khoschy; Becker, Anton S.; Donati, Olivio; Konukoglu, Ender. PHiSeg: Capturing Uncertainty in Medical Image Segmentation. Medical Image Computing and Computer Assisted Intervention (MICCAI). arXiv:1906.04045, 2019.

Bouchard, Mathieu; Jousselme, Anne-Laure; Dore, Pierre-Emmanuel. A Proof for the Positive Definiteness of the Jaccard Index Matrix. International Journal of Approximate Reasoning, 54:615–626. DOI: 10.1016/j.ijar.2013.01.006, 2013.

Broder, Andrei Z.. On the Resemblance and Containment of Documents. Proceedings of the Compression and Complexity of Sequences (SEQUENCES):21–29. DOI: 10.1109/SEQUEN.1997.666900, 1997.

Cordts, Marius; Omran, Mohamed; Ramos, Sebastian; Rehfeld, Timo; Enzweiler, Markus; Benenson, Rodrigo; Franke, Uwe; Roth, Stefan; Schiele, Bernt. The Cityscapes Dataset for Semantic Urban Scene Understanding. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

Daras, Giannis; Delbracio, Mauricio; Talebi, Hossein; Dimakis, Alexandros G.; Milanfar, Peyman. Soft Diffusion: Score Matching for General Corruptions. arXiv preprint arXiv:2209.05442, 2022.

De Sousa Ribeiro, Fabio; Todd, Omar; Jones, Charles; Kori, Avinash; Mehta, Raghav; Glocker, Ben. Flow Stochastic Segmentation Networks. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2025.

Gao, Zhitong; Chen, Yucong; Zhang, Chuyu; He, Xuming. Modeling Multimodal Aleatoric Uncertainty in Segmentation with Mixture of Stochastic Experts. International Conference on Learning Representations (ICLR), 2023.

Gneiting, Tilmann; Raftery, Adrian E.. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association, 102:359–378. DOI: 10.1198/016214506000001437, 2007.

Gower, John C.; Legendre, Pierre. Metric and Euclidean Properties of Dissimilarity Coefficients. Journal of Classification, 3:5–48. DOI: 10.1007/BF01896809, 1986.

Gretton, Arthur; Borgwardt, Karsten M.; Rasch, Malte J.; Scholkopf, Bernhard; Smola, Alexander. A Kernel Two-Sample Test. Journal of Machine Learning Research, 13:723–773, 2012.

Guo, Chuan; Pleiss, Geoff; Sun, Yu; Weinberger, Kilian Q.. On Calibration of Modern Neural Networks. International Conference on Machine Learning (ICML), 2017.

Hoogeboom, Emiel; Salimans, Tim. Blurring Diffusion Models. International Conference on Learning Representations (ICLR), 2023.

Jurgens, Mira; Meinert, Nis; Bengs, Viktor; Hullermeier, Eyke; Waegeman, Willem. Is Epistemic Uncertainty Faithfully Represented by Evidential Deep Learning Methods?. International Conference on Machine Learning (ICML). arXiv:2402.09056, 2024.

Kahl, Kim-Celine; Luth, Carsten T.; Zenk, Maximilian; Maier-Hein, Klaus; Jaeger, Paul F.. ValUES: A Framework for Systematic Validation of Uncertainty Estimation in Semantic Segmentation. International Conference on Learning Representations (ICLR). arXiv:2401.08501, 2024.

Karandikar, Archit; Cain, Nicholas; Tran, Dustin; Lakshminarayanan, Balaji; Shlens, Jonathon; Mozer, Michael C.; Roelofs, Rebecca. Soft Calibration Objectives for Neural Networks. Advances in Neural Information Processing Systems (NeurIPS), 2021.

Kassapis, Elias; Dikov, Georgi; Gupta, Deepak K.; Nugteren, Cedric. Calibrated Adversarial Refinement for Stochastic Semantic Segmentation. IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

Kerrigan, Gavin; Migliorini, Giosue; Smyth, Padhraic. Functional Flow Matching. International Conference on Artificial Intelligence and Statistics (AISTATS), 2024.

Kohl, Simon A. A.; Romera-Paredes, Bernardino; Meyer, Clemens; De Fauw, Jeffrey; Ledsam, Joseph R.; Maier-Hein, Klaus H.; Eslami, S. M. Ali; Rezende, Danilo Jimenez; Ronneberger, Olaf. A Probabilistic U-Net for Segmentation of Ambiguous Images. Advances in Neural Information Processing Systems (NeurIPS), 2018.

Kohl, Simon A. A.; Romera-Paredes, Bernardino; Maier-Hein, Klaus H.; Rezende, Danilo Jimenez; Eslami, S. M. Ali; Kohli, Pushmeet; Zisserman, Andrew; Ronneberger, Olaf. A Hierarchical Probabilistic U-Net for Modeling Multi-Scale Ambiguities. arXiv preprint arXiv:1905.13077, 2019.

Li, Fanding; Li, Xiangyu; Su, Xianghe; Qiu, Xingyu; Dong, Suyu; Wang, Wei; Wang, Kuanquan; Luo, Gongning; Li, Shuo. Ambiguity-aware Truncated Flow Matching for Ambiguous Medical Image Segmentation. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2026.

Lipman, Yaron; Chen, Ricky T. Q.; Ben-Hamu, Heli; Nickel, Maximilian; Le, Matt. Flow Matching for Generative Modeling. International Conference on Learning Representations (ICLR), 2023.

Lyons, Russell. Distance Covariance in Metric Spaces. The Annals of Probability, 41:3284–3305. DOI: 10.1214/12-AOP803, 2013.

Monteiro, Miguel; Le Folgoc, Loic; Coelho de Castro, Daniel; Pawlowski, Nick; Marques, Bernardo; Kamnitsas, Konstantinos; van der Wilk, Mark; Glocker, Ben. Stochastic Segmentation Networks: Modelling Spatially Correlated Aleatoric Uncertainty. Advances in Neural Information Processing Systems (NeurIPS), 2020.

Pedrosa, Joao; Aresta, Guilherme; Ferreira, Carlos; Rodrigues, Marcio; Leitao, Patricia; Carvalho, Andre Silva; Rebelo, Joao; Negrao, Eduardo; Ramos, Isabel; Cunha, Antonio; Campilho, Aurelio. LNDb: A Lung Nodule Database on Computed Tomography. arXiv preprint arXiv:1911.08434, 2019.

Rahman, Aimon; Valanarasu, Jeya Maria Jose; Hacihaliloglu, Ilker; Patel, Vishal M.. Ambiguous Medical Image Segmentation Using Diffusion Models. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

Rissanen, Severi; Heinonen, Markus; Solin, Arno. Generative Modelling with Inverse Heat Dissipation. International Conference on Learning Representations (ICLR), 2023.

Sejdinovic, Dino; Sriperumbudur, Bharath; Gretton, Arthur; Fukumizu, Kenji. Equivalence of Distance-Based and RKHS-Based Statistics in Hypothesis Testing. The Annals of Statistics, 41:2263–2291. DOI: 10.1214/13-AOS1140, 2013.

Sensoy, Murat; Kaplan, Lance; Kandemir, Melih. Evidential Deep Learning to Quantify Classification Uncertainty. Advances in Neural Information Processing Systems (NeurIPS), 2018.

Szekely, Gabor J.; Rizzo, Maria L.. Energy Statistics: A Class of Statistics Based on Distances. Journal of Statistical Planning and Inference, 143:1249–1272. DOI: 10.1016/j.jspi.2013.03.018, 2013.

Xie, Enze; Wang, Wenhai; Yu, Zhiding; Anandkumar, Anima; Alvarez, Jose M.; Luo, Ping. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. Advances in Neural Information Processing Systems (NeurIPS), 2021.

Zbinden, Lukas; Doorenbos, Lars; Pissas, Theodoros; Huber, Adrian Thomas; Sznitman, Raphael; Marquez-Neila, Pablo. Stochastic Segmentation with Conditional Categorical Diffusion Models. International Conference on Computer Vision (ICCV). arXiv:2303.08888, 2023.

A Metric definitions and a complete finite-domain proof

Lemma 1 (Strict positive definiteness of Jaccard on a finite grid). For a finite set Ω, the matrix $K_{AB} = \mathrm{IoU}(A, B)$ indexed by all subsets of Ω, with the stated empty-mask convention, is strictly positive definite.

Proof.

For a uniformly random permutation π of Ω, let $h_\pi(A)$ be the first element of a nonempty A and let $h_\pi(\varnothing) = \perp$. The MinHash identity (Broder, 1997) gives $K_{AB} = \mathbb{P}(h_\pi(A) = h_\pi(B))$. Consequently,

$$v^\top K v = \mathbb{E}_\pi \sum_b \Big(\sum_{A\,:\,h_\pi(A)=b} v_A \Big)^2 \geq 0. \tag{11}$$

If the expression is zero, every bucket sum is zero for every permutation, since the permutation space is finite and each permutation has positive probability. The special bucket gives $v_\varnothing = 0$. Fix $b \in \Omega$ and any $T \subseteq \Omega \setminus \{b\}$. Choose a permutation placing all elements outside $T \cup \{b\}$ before b, and every element of T after b. Its bucket for b contains precisely the sets $\{b\} \cup U$ for $U \subseteq T$, so

$$\sum_{U \subseteq T} v_{\{b\} \cup U} = 0. \tag{12}$$

Taking $T = \varnothing$ gives $v_{\{b\}} = 0$. Induction on $|T|$ then gives $v_{\{b\} \cup T} = 0$ for every T. Every nonempty mask contains a possible b, so all coefficients vanish. Principal submatrices inherit strict positive definiteness. This completes the finite-domain argument, building on the strict-PD result of Bouchard et al. (2013). The implication from strict positive definiteness to characteristicness is used here on a finite mask space, not on an arbitrary infinite domain.

End of proof.

Empirical estimators. Writing the finite-sample averages explicitly,

$$C = \tfrac{1}{KR} \sum_{i=1}^{K} \sum_{j=1}^{R} d(s_i, y_j), \qquad W_S^U = \tfrac{1}{K(K-1)} \sum_{i \neq j} d(s_i, s_j), \tag{13}$$

with the analogous definition for raters. Since $W_S^V = (K-1) W_S^U / K$ and $W_Y^V = (R-1) W_Y^U / R$, Eq. 4 follows exactly for any distance with zero diagonal. The feature-mean identity holds for the population divergence and for the empirical V-statistic; the corresponding U-statistic is the unbiased kernel U estimator, not the squared norm of empirical means.

A diagnostic counterexample. Take identical sample and rater sets containing two distinct disjoint nonempty masks A, B. For either ground distance used here, $C = 1/2$ and $W_S^U = W_Y^U = 1$, so $\widehat{G}_U = -1$ while $\widehat{G}_V = 0$. This does not contradict nonnegative population GED. Checks must separately cover these estimator behaviors, exclude self-pairs by index, retain duplicate draws, and avoid clipping negative U estimates. The example is a useful regression test for distinguishing the two estimators.

HM-IoU conventions. For LIDC, four annotations tile evenly to $K = 4$ or 16 before a square assignment. Rectangular $K \times R$ matching instead scores only a selected subset of generated masks and can inflate apparent quality. For rater counts that do not divide 16, cyclic tiling gives unequal rater weights; a multiple-of-R sample count or an explicitly weighted transport formulation avoids that imbalance.

Table 4. Internal-LIDC HM-IoU under tiled and rectangular matching. Most methods have nine seeds; SSN rank 50 has three. The $K = 4$ and $K = 16$ columns show sensitivity to the number of samples.

Method	tiled $K = 4$	tiled $K = 16$	rect. $K \times R$ ($K = 16$)
Deterministic	0.462±0.009	0.462±0.009	0.462±0.009
MC-Dropout	0.492±0.004	0.495±0.005	0.571±0.014
Evidential	0.285±0.012	0.287±0.012	0.301±0.011
Prob. U-Net ($b = 1$, collapsed)	0.461±0.009	0.462±0.009	0.467±0.008
Prob. U-Net ($b = 0.1$, tuned)	0.567±0.013	0.598±0.012	0.815±0.002
SSN (default)	0.478±0.015	0.480±0.015	0.493±0.012
SSN (covariance-tempered)	0.398±0.070	0.404±0.071	0.491±0.089
SSN (rank 50)	0.471±0.020	0.472±0.019	0.487±0.019
i.i.d.-noise ablation	0.351±0.026	0.361±0.027	0.476±0.028
EviFlow (ours)	0.470±0.028	0.489±0.029	0.671±0.033

Table 5. Internal-LIDC V-statistic GED at $K = 16$, for comparison with the U-statistic values in Table 1. Comparisons across publications require matched preprocessing, sample counts, ground distances, and matching conventions. SSN rank 50 has three seeds; the other rows have nine.

Method	GED_V ($\beta = 2$)	GED_V ($\beta = 1$)
Deterministic	0.662±0.018	0.819±0.027
MC-Dropout	0.515±0.017	0.554±0.023
Evidential	0.435±0.003	0.447±0.003
Prob. U-Net ($b = 1$, collapsed)	0.642±0.017	0.735±0.020
Prob. U-Net ($b = 0.1$, tuned)	0.250±0.010	0.281±0.008
SSN (default)	0.556±0.018	0.587±0.029
SSN (covariance-tempered)	0.428±0.016	0.433±0.017
SSN (rank 50)	0.524±0.024	0.542±0.038
i.i.d.-noise ablation	0.373±0.016	0.388±0.015
EviFlow (ours)	0.279±0.022	0.295±0.023

B Datasets and LNDb evaluation

The internal LIDC pipeline contains 445 patients, 1,388 annotated nodules, and 8,262 axial crops, split into 5,967/682/1,613 train/validation/test crops from 315/40/90 patients. Each crop stores four rater slots, with absent annotations represented by empty masks. Raw contour counts exceed four for 82 crops because clustering or repeated outlines can produce additional contours; the four-primary-slot convention is retained. Table 6 gives the contour-count distribution in the patient-disjoint splits.

Table 6. Raw contour counts in the internal LIDC split.

Split	1	2	3	4	5	7	Total
Train	1564	978	1103	2273	28	21	5967
Validation	139	104	114	325	0	0	682
Test	356	268	403	553	33	0	1613

The LNDb experiment (Pedrosa et al., 2019) uses three annotation slots and 1,473 crops split at patient level into 1,014/139/320 train/validation/test crops. This is a separately trained cohort, not the LIDC-to-LNDb transfer experiment used for risk–coverage evaluation. The KL-weight selection is dataset-specific.

Table 7. LNDb U-statistic results, nine seeds. The paired energy-GED test against the Probabilistic U-Net gives unadjusted $p = 0.020$; against the i.i.d.-noise ablation it gives $p = 0.098$. The HM-IoU difference from the Probabilistic U-Net is not significant.

Method	$D^2_{\mathrm{GED},\, r=1/2}$	$D^2_{\mathrm{GED},\, r=1}$	HM-IoU	disp. ratio$_U$
Prob. U-Net (tuned)	0.076±0.028	0.095±0.033	0.552±0.022	0.57±0.07
i.i.d.-noise ablation	0.114±0.116	0.136±0.130	0.506±0.075	0.68±0.47
EviFlow (ours)	0.033±0.046	0.050±0.054	0.563±0.074	0.80±0.24

C Training, auxiliary heads, and calibration

The LIDC recipe uses SegFormer-B2, 100 epochs of AdamW, learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 32, bfloat16, and gradient clipping at 1. The checkpoint rule is the lowest validation value of each method's own training objective. The conditioning width is 32, blur size 17, and Euler step count 8. Loss weights are $\lambda_{\mathrm{flow}} = \lambda_{\mathrm{dice}} = 1$, $\lambda_{\mathrm{ece}} = 0.5$, and $\lambda_{\mathrm{cons}} = 0.1$. The i.i.d.-noise ablation additionally uses $\lambda_{\mathrm{div}} = 0.1$; the evidential KL term is annealed over the first 10 epochs. The base field is Gaussian-derived: per-sample standard-deviation normalization changes its joint Gaussian law while preserving spatial correlation as the intended inductive bias.

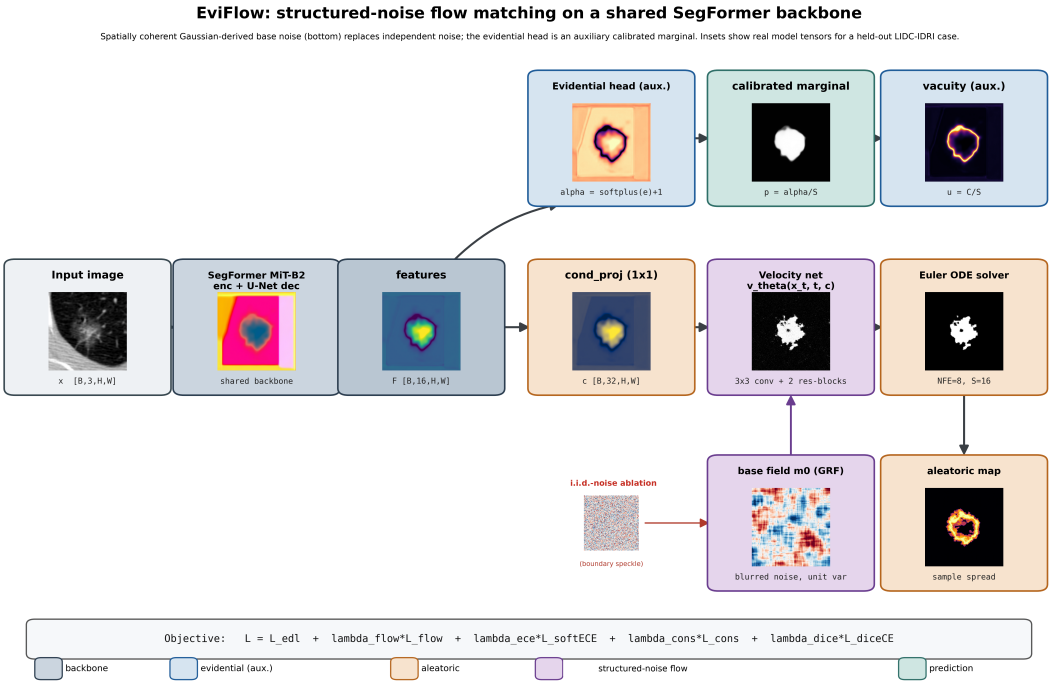

Figure 3. EviFlow architecture with example tensors. The evidential head provides an auxiliary marginal; the conditioned velocity network transforms a correlated base field into samples. The structured base field is used in both training and inference.

Table 8. Probabilistic U-Net validation sweep. Here b is the KL coefficient, distinct from GED's energy-family β. The selected setting is $b = 0.1$. The table shows validation HM-IoU and energy-distance dispersion, with nine seeds for each coefficient.

b (KL)	Seeds	Validation HM-IoU	Validation ρ_U
0.01	9	0.476	1.18
0.03	9	0.502	1.15
0.1	9	0.575	1.03 $\leftarrow$ selected
0.3	9	0.484	0.20
1	9	0.487	0.16
3	9	0.486	0.18
10	9	0.489	0.19

Evidential versus softmax. The exploratory softmax-head comparison gives similar HM-IoU means: 0.488 for softmax and 0.489 for the evidential variant. Both heads can provide calibrated per-pixel probabilities. We retain the evidential head as an auxiliary marginal predictor and evaluate its additional vacuity signal separately in Table 9.

Table 9. Risk–coverage AURC, averaged over three seeds; lower is better. Risk is one minus IoU to rater consensus. LIDC-to-LNDb denotes transfer, not separately trained LNDb models.

Uncertainty signal	AURC (LIDC)	AURC (LIDC $\rightarrow$ LNDb)
EviFlow evidential vacuity	0.438	0.396
EviFlow marginal entropy	0.429	0.391
MC-Dropout entropy	0.432	0.365
(random ordering)	0.465	0.448

Table 10. Per-pixel uncertainty metrics on internal LIDC, nine seeds. ECE and AUSE are lower-is-better; disagreement AUROC is higher-is-better.

Method	ECE_{band}	disagree AUROC	AUSE
Deterministic	0.096±0.003	0.874±0.011	0.007±0.008
MC-Dropout	0.088±0.003	0.884±0.003	0.005±0.001
Evidential	0.125±0.003	0.888±0.005	0.008±0.001
Prob. U-Net ($b = 0.1$, tuned)	0.119±0.004	0.878±0.004	0.005±0.000
SSN	0.099±0.002	**0.890±0.004**	**0.005±0.000**
i.i.d.-noise ablation	**0.085±0.006**	0.867±0.011	0.062±0.031
EviFlow (ours)	0.090±0.009	0.855±0.021	0.101±0.074

D Additional sample-fidelity experiments

We tested inference-time base scaling and 8, 16, or 32 Euler steps; these settings did not close the HM-IoU gap to the selected Probabilistic U-Net. We also tested a mean-anchored bridge initialized at $a = 4(2\bar{p} - 1)$ plus correlated noise. The best validation HM-IoU is 0.404±0.076 for that bridge, versus 0.506±0.028 for unanchored EviFlow. Among these settings, the unanchored structured-noise sampler gives the stronger sample-to-rater correspondence.

E Cross-dataset evaluation

We evaluate complementary regimes on canonical LIDC, the nine QUBIQ tasks, and stochastic Cityscapes. The protocols differ in rater counts, class structure, and evaluation conventions; Table 16 summarizes the resulting trade-offs within each benchmark.

Canonical LIDC cohort. The Kohl split has 1,980 test crops and four annotation slots. The comparison contains three seeds for the matched-backbone methods and one for Flow-SSN. Flow-SSN is a separately implemented reference, not a matched-backbone variant. Flow-SSN has the lowest mean GED, while the Probabilistic U-Net has the highest mean HM-IoU; EviFlow is third on both metrics. This split illustrates why distributional fit and matched sample fidelity should be reported together.

Table 11. Canonical LIDC comparison. Values are mean± seed SD where $n > 1$; the single Flow-SSN run has no estimated seed SD. GED uses U-statistics, HM-IoU tiled $K = 16$ matching, and dispersion the energy distance. Dashes indicate unreported metrics.

Method	n	$\mathrm{GED}_U\ \beta = 1$	$\mathrm{GED}_U\ \beta = 2$	HM-IoU	ρ_U
Deterministic	3	–	–	0.443±0.008	–
MC-Dropout	3	0.468±0.069	0.460±0.059	0.466±0.014	0.409±0.098
Evidential	3	0.265±0.004	0.291±0.004	0.298±0.008	1.256±0.017
SSN	3	0.468±0.050	0.470±0.034	0.463±0.002	0.344±0.102
Prob. U-Net ($b = 0.1$)	3	0.144±0.004	0.141±0.004	0.577±0.006	0.922±0.029
Flow-SSN	1	0.093	0.108	0.531	1.138
EviFlow	3	0.168±0.104	0.196±0.123	0.511±0.049	0.868±0.248

QUBIQ: all nine tasks. QUBIQ-2021 comprises nine binary tasks with different organs, rater counts, and small test cohorts. Each task uses three seeds per method, with SegFormer-B2, AdamW, 150 epochs, and batch size 16. Table 12 reports an equal-weight macro-average of the nine task means for the same reference configurations shown in the per-task tables. EviFlow has the lowest macro GED under both ground distances and the highest macro HM-IoU among the stochastic methods. The Deterministic model has the highest overall macro HM-IoU, again exposing a distinction between distributional comparison and matched fidelity. Tables 13 and 14 retain all nine tasks, including tasks with weak overlap, to make this heterogeneity visible.

Table 12. QUBIQ reference configurations: equal-weight macro means over all nine tasks, with three seeds per task and method. These reference settings are not asserted to be validation-optimal. Per-task seed SDs appear in Tables 13–14; the macro table reports means only. Dashes indicate unreported metrics.

Method	$\mathrm{GED}_U\ \beta = 1$	$\mathrm{GED}_U\ \beta = 2$	$\mathrm{HM\text{-}IoU}_{16}$	ρ_U
Deterministic	–	–	0.409	–
MC-Dropout	0.527	0.639	0.344	1.432
Evidential	0.672	0.950	0.058	2.115
SSN	0.601	0.824	0.104	2.152
Prob. U-Net	0.970	1.103	0.276	0.397
EviFlow	0.338	0.424	0.389	1.777

Table 13. QUBIQ reference cohort, per-task energy GED_U: mean±SD over three seeds. Rater counts are shown in parentheses. Lower is better; dashes denote unreported metrics.

Task (R)	Determ.	MC-Drop	Evid.	SSN	PUNet	EviFlow
Brain growth t1 (7)	–	0.277±0.012	0.530±0.028	0.434±0.071	0.877±0.633	0.123±0.028
Brain tumor t1 (3)	–	0.330±0.054	0.760±0.025	0.677±0.054	0.982±0.694	0.141±0.015
Brain tumor t2 (3)	–	0.316±0.083	0.432±0.037	0.410±0.010	1.052±0.408	0.197±0.048
Brain tumor t3 (3)	–	0.640±0.076	0.605±0.036	0.586±0.018	1.324±0.051	0.584±0.167
Kidney t1 (3)	–	0.981±0.275	0.867±0.030	0.872±0.060	1.689±0.011	0.630±0.114
Pancreas t1 (2)	–	0.681±0.182	0.616±0.029	0.608±0.017	1.209±0.423	0.490±0.046
Pancreatic lesion t1 (2)	–	1.047±0.254	0.944±0.024	0.944±0.044	0.956±0.692	0.587±0.207
Prostate t1 (6)	–	0.206±0.026	0.663±0.033	0.420±0.090	0.247±0.076	0.073±0.008
Prostate t2 (6)	–	0.261±0.011	0.629±0.029	0.454±0.059	0.397±0.238	0.213±0.155

Table 14. QUBIQ reference cohort, per-task tiled HM-IoU at $K = 16$: mean±SD over three seeds. Higher is better. The tiling imbalance for rater counts not dividing 16 is discussed in Appendix A.

Task (R)	Determ.	MC-Drop	Evid.	SSN	PUNet	EviFlow
Brain growth t1 (7)	0.718±0.015	0.665±0.025	0.145±0.010	0.210±0.045	0.282±0.396	0.671±0.016
Brain tumor t1 (3)	0.764±0.042	0.679±0.056	0.064±0.006	0.091±0.029	0.282±0.444	0.779±0.022
Brain tumor t2 (3)	0.422±0.052	0.130±0.123	0.023±0.007	0.020±0.003	0.010±0.017	0.226±0.080
Brain tumor t3 (3)	0.084±0.075	0.023±0.024	0.006±0.001	0.005±0.000	0.037±0.028	0.024±0.031
Kidney t1 (3)	0.012±0.007	0.014±0.006	0.020±0.001	0.019±0.001	0.007±0.011	0.135±0.232
Pancreas t1 (2)	0.020±0.012	0.009±0.001	0.009±0.001	0.008±0.000	0.002±0.004	0.002±0.003
Pancreatic lesion t1 (2)	0.000±0.000	0.001±0.000	0.001±0.000	0.001±0.000	0.400±0.346	0.117±0.151
Prostate t1 (6)	0.865±0.014	0.822±0.051	0.158±0.009	0.367±0.089	0.878±0.016	0.854±0.012
Prostate t2 (6)	0.797±0.003	0.751±0.054	0.096±0.007	0.219±0.051	0.590±0.373	0.697±0.115

Official stochastic Cityscapes score. The protocol uses 256-by-512 images and the five class-switch construction of Kohl et al. (2018). Its official metric compares samples with the probability-weighted 32-mode reference over ten switchable classes, omitting jointly absent classes from each class average. The official-score comparison uses $K = 16$, nine seeds for EviFlow, and three for each baseline. EviFlow has the lowest observed mean GED in this comparison, 0.351 versus MC-Dropout's 0.380. We report the mean comparison with its seed variability.

Table 15. Official stochastic Cityscapes metric at $K = 16$: mean±seed SD, with seed counts shown. Probabilistic U-Net is a reference configuration, not the LIDC-tuned result. The metric is distinct from binary and class-averaged surrogate GEDs.

Method	n	Official GED
Deterministic	3	0.754±0.004
MC-Dropout	3	0.380±0.021
Evidential	3	0.463±0.037
SSN	3	0.450±0.003
Prob. U-Net	3	0.624±0.009
EviFlow	9	0.351±0.039

For context, Kassapis et al. (2021) report GED 0.206 for the Probabilistic U-Net and 0.164 for CAR in their Cityscapes setting. The gap to the present reference row motivates further baseline

optimization; the different pipelines make these literature values contextual rather than like-for-like comparisons.

Why the binary theorem does not settle multiclass evaluation. For a fixed class-averaged IoU kernel, the mean embedding sees each class's marginal mask distribution, not every cross-class dependency. With two pixels and four classes, $P = \frac{1}{2}(a, c) + \frac{1}{2}(b, d)$ and $Q = \frac{1}{2}(a, d) + \frac{1}{2}(b, c)$ have identical per-class marginals but different joint laws. Their class-averaged-kernel GED is zero. Thus that fixed-kernel divergence is not strictly identifying. The official absent-class nanmean is pair-dependent and is not the same fixed kernel; the binary proof does not establish its propriety.

Table 16. Cross-dataset overview. "Lowest" and "highest" refer to observed means, not significance. QUBIQ includes all nine tasks. Metrics and seed counts differ across rows, so absolute GED values are not comparable between datasets.

Dataset / cohort	Lowest observed GED	Highest observed HM-IoU
LIDC-kohl (Flow $n = 1$; others 3)	Flow-SSN (0.093)	Prob. U-Net (0.577)
QUBIQ (all nine tasks; $n = 3$/task)	EviFlow (0.338)	Deterministic (0.409)
Cityscapes (EviFlow $n = 9$; baselines 3)	EviFlow (0.351)	–

Across these benchmarks, no single method leads on every criterion. The combined picture supports evaluating complete mask distributions and their individual samples jointly, with each comparison tied to an explicit dataset and metric protocol.