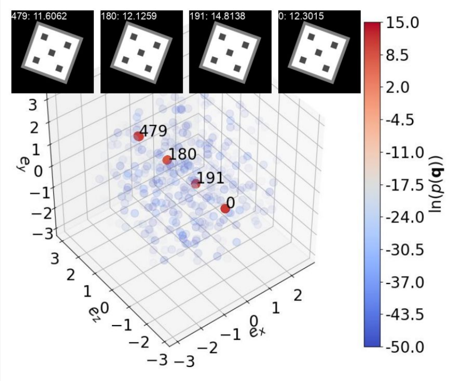
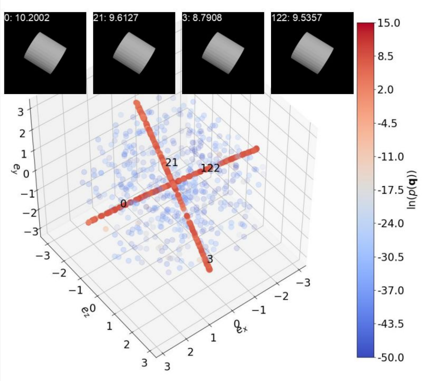
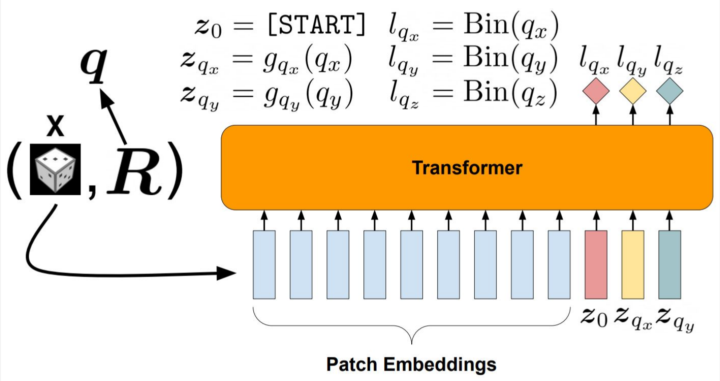
AQuaMaM: An Autoregressive, Quaternion Manifold Model for Rapidly Estimating Complex **SO**(3) Distributions

Michael A. Alcorn







Modeling Uncertainty in 3D Rotations is a Fundamental Bottleneck in Robotics

- Estimating the pose of objects is a prerequisite for many robotics applications, from manipulation to navigation.
- This is uniquely challenging because the 3D rotation group, **SO**(3), lies on a *curved* manifold, making standard probability distributions like the multivariate Gaussian unsuitable.
- Crucially, models must account for multimodality.

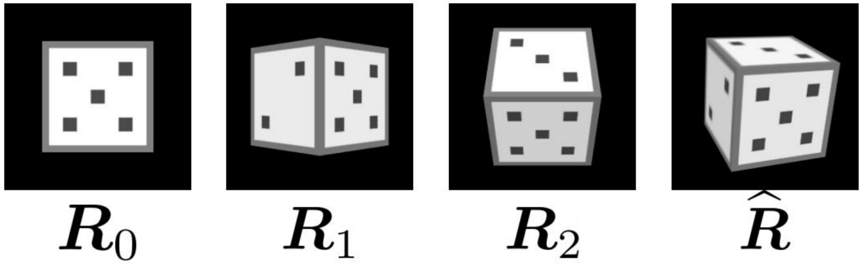


Figure 1: When minimizing the unimodal Bingham loss for the two rotations $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$, the maximum likelihood estimate $\widehat{\boldsymbol{R}}$ is a rotation that was never observed in the dataset. Note, the die images are for demonstration purposes only, i.e., no images were used during optimization. $\boldsymbol{R}_0$ is the identity rotation.

IPDF Has A Trade-off Between Precision and Speed

- Implicit-PDF (IPDF) is an elegant and effective approach for modeling distributions on **SO**(3).
- **Its Bottleneck**: Inference requires *N* forward passes through its network to calculate likelihood, where *N* determines the model's precision.
  - This is prohibitively slow without massive parallelization.
- **A Hidden Problem**: IPDF is typically trained with a much smaller *N* than is used for testing (e.g., train = 4,096 vs test = 2,359,296).
  - Makes it difficult to reason about how the model will behave in the wild.

$$p(R|x) \approx \frac{1}{V} \frac{\exp(f(x, R))}{\sum_i^N \exp(f(x, R_i))},$$

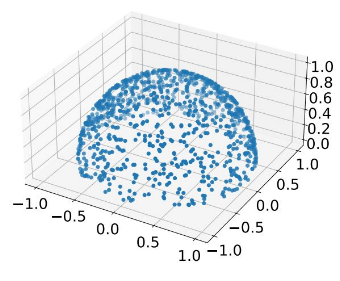
$$V = \pi^2 / N$$

Volume of **SO**(3)

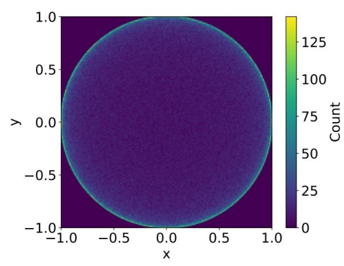
Reframe the Problem: Instead of Modeling a Curved Manifold, Model Its Flat Projection

- Directly modeling distributions on the curved 3-sphere of unit quaternions is difficult.
- **The Key Insight**: We can uniquely represent every possible rotation using only the first three components of a unit quaternion (q_x, q_y, q_z).
  - These three components must lie within a standard, non-curved unit 3-ball (B^3).
- This creates a bijective mapping from a simple, flat space (the 3-ball) to the complex, curved space of rotations (the “hyper-hemisphere” ~H^1).
- We can now model the distribution in this simpler space and use [a density transformation]() to get the exact probability on the manifold.

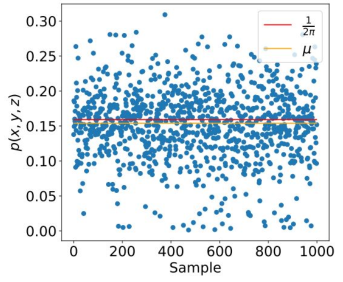
Samples from ~S^2



Projected onto B^2 and binned

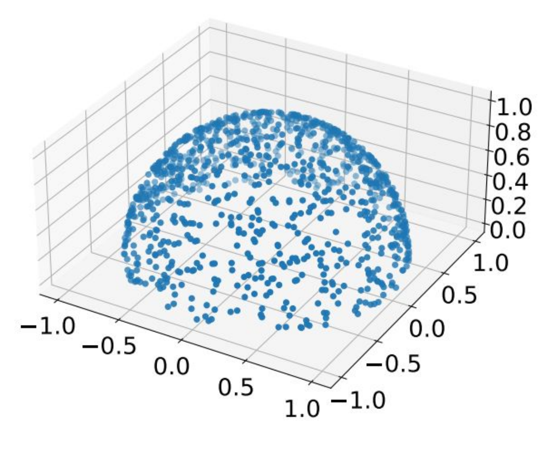


Estimated densities

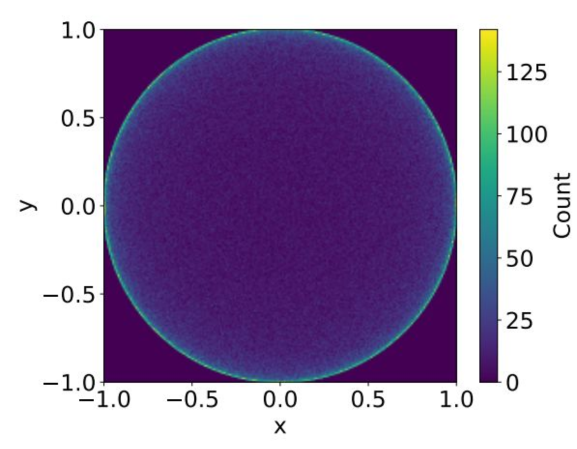


AQuaMaM: A “Quaternion Language Model” for rotations

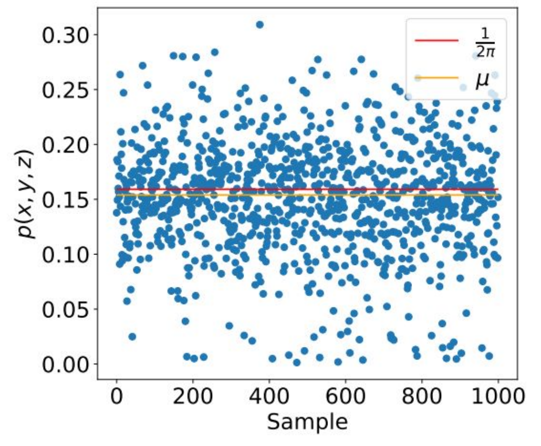
Samples from ~S^2



Projected onto B^2 and binned



Estimated densities



$$f(x, y) = [x, y, z]$$

$$z = \sqrt{1 - x^2 - y^2}$$

$$\boldsymbol{J} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{-x}{z} & \frac{-y}{z} \end{bmatrix}$$

$$a = \sqrt{\begin{vmatrix} 0 & 1 \\ \frac{-x}{z} & \frac{-y}{z} \end{vmatrix}^2 + \begin{vmatrix} 1 & 0 \\ \frac{-x}{z} & \frac{-y}{z} \end{vmatrix}^2 + \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}^2}$$

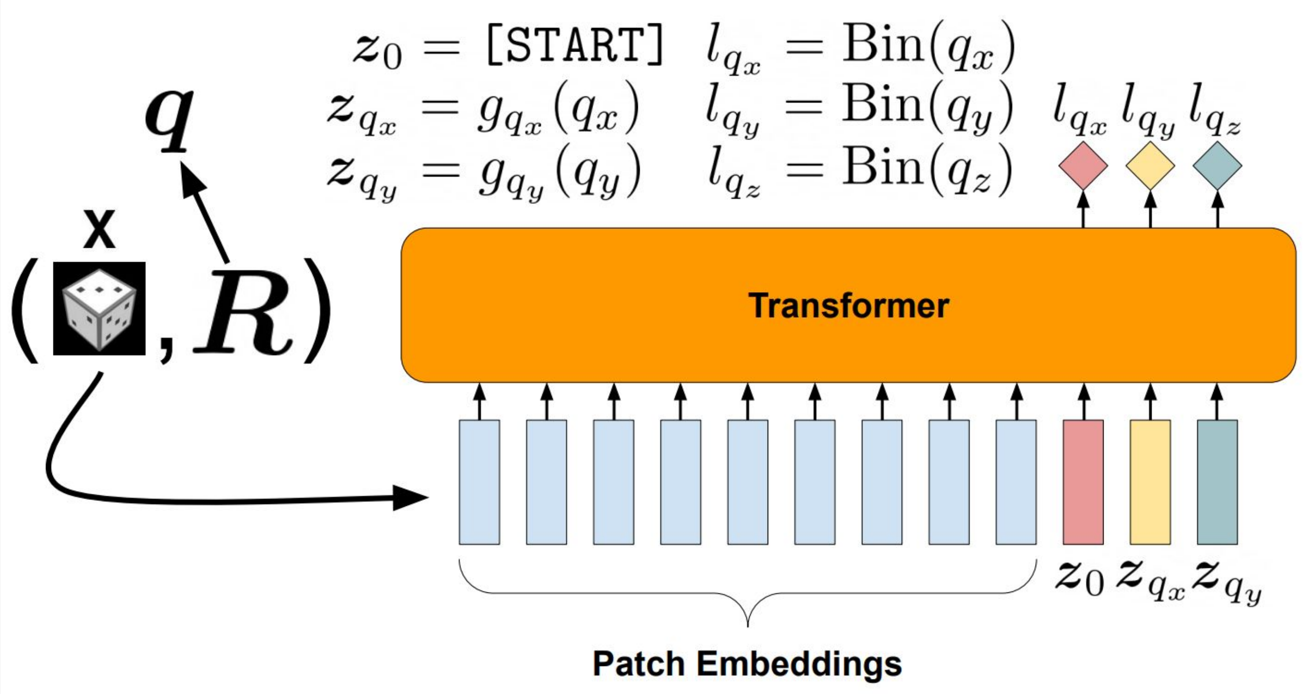
$$= \sqrt{\frac{x^2}{z^2} + \frac{y^2}{z^2} + 1} = \sqrt{\frac{x^2 + y^2 + z^2}{z^2}} = \frac{1}{z}$$

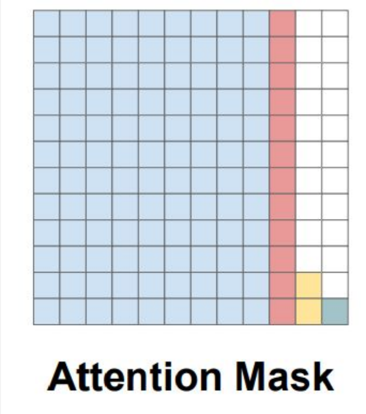
$$p(x, y, z) = \frac{p(x,y)}{a}$$

$$= p(x, y)z$$

$$= p(x)p(y|x)z$$

The Architecture is an Extended Vision Transformer with a Partially Causal Attention Mask





$$p(q_x, q_y, q_z) = \pi_{q_x} \frac{N}{2} \pi_{q_y} \frac{1}{\omega_{q_y}} \pi_{q_z} \frac{1}{\omega_{q_z}}$$

$$= \pi_{q_x} \pi_{q_y} \pi_{q_z} \frac{N}{2\omega_{q_y}\omega_{q_z}}$$

$$\mathcal{L} = -\sum_{d=1}^{|\mathcal{X}|} \ln \pi_{q_{d,x}} + \ln \pi_{q_{d,y}} + \ln \pi_{q_{d,z}} + \ln \frac{N q_{d,w}}{2\omega_{q_{d,y}}\omega_{q_{d,z}}}$$

Constant

AQuaMaM Intelligently Enforces Constraints

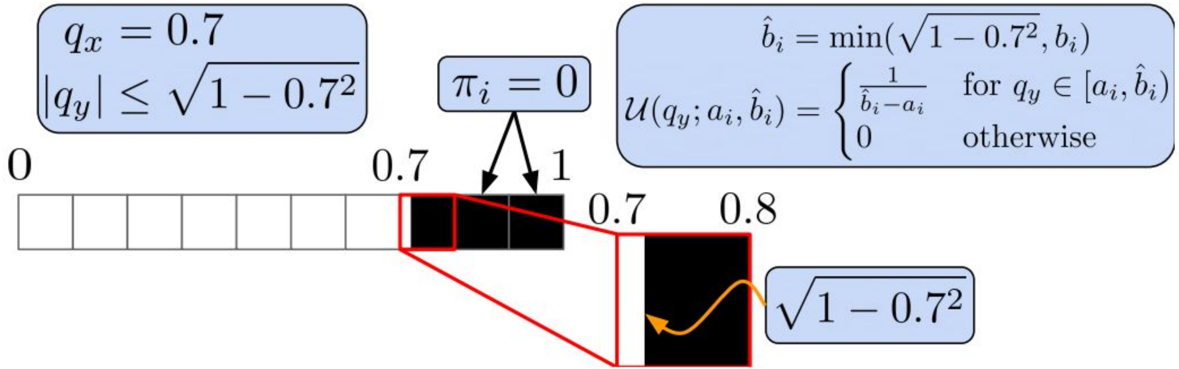


Figure 3: When modeling the conditional distribution $p(q_y|q_x)$ as a mixture of uniform distributions, the geometric constraints of the unit quaternion are easily enforced. Here, I focus on non-negative bins for clarity, i.e., intervals $[a_i, b_i)$ where $0 \leq a < b \leq 1$, but the same logic applies to negative bins. Given $q_x = 0.7$, we know that $|q_y| \leq \sqrt{1 - 0.7^2}$ because $\boldsymbol{q}$ has a unit norm. As a result, the mixture proportion π_i for any bin where $\sqrt{1 - 0.7^2} < a_i$ *must* be zero. AQuaMaM enforces this constraint by assigning a value of $-\infty$ to the output scores for "strictly illegal bins" during training.[10] For the remaining bins, the corresponding uniform distribution is $\mathcal{U}(q_y; a_i, \hat{b}_i)$ where $\hat{b}_i = \min(\sqrt{1 - 0.7^2}, b_i)$, i.e., the upper bound of the uniform distribution for the partially legal bin is reduced to $\sqrt{1 - 0.7^2}$.

On a Synthetic Dataset, AQuaMaM Recovers the True Data Distribution While IPDF Dramatically Diverges

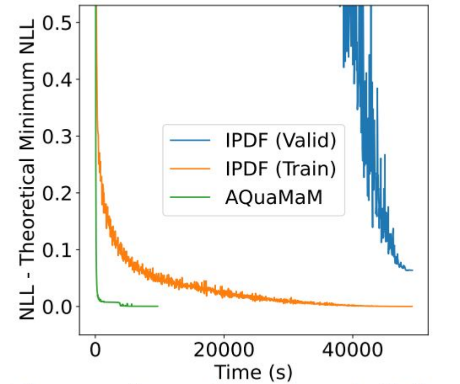
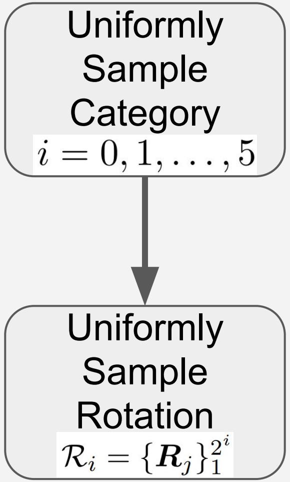


Figure 5: On the infinite toy dataset, AQuaMaM rapidly reached its theoretical minimum (classification) average negative log-likelihood (NLL). In contrast, IPDF never reached its theoretical minimum validation NLL, despite converging to its training theoretical minimum.



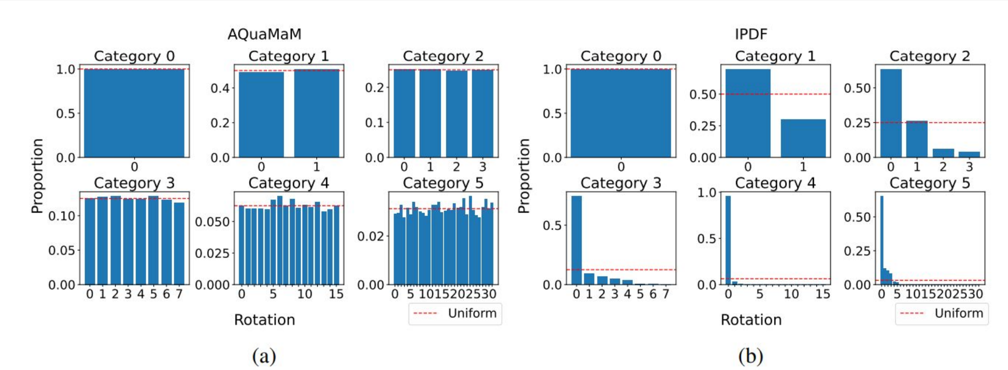


Figure 6: (a) The proportions of sampled rotations from the AQuaMaM model trained on the infinite toy dataset closely approximate the expected uniform distributions. (b) In contrast, despite approaching its theoretical minimum log-likelihood during training (Figure 5), the proportions of sampled rotations from the IPDF model drastically diverge from the expected uniform distributions.

Model	**Average LL (↑)**	**Average Distance (↓)**
IPDF	12.32	0.84°
AQuaMaM	**27.12**	**0.04°**

IPDF would need to use six *trillion* cells for it to be theoretically possible to match AQuaMaM's average log-likelihood

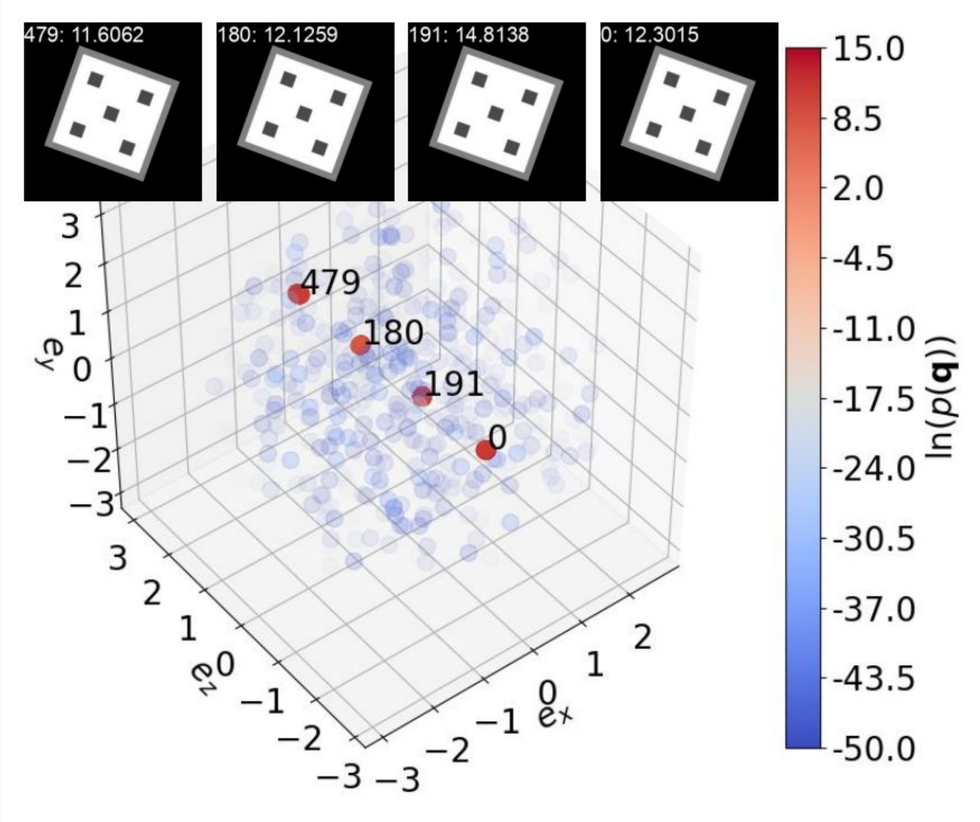
$$\frac{N q_w}{2\omega_{q_y}\omega_{q_z}} \geq \frac{N^3 q_w}{8}$$

On a 500,000-Image Die Dataset, AQuaMaM Achieves Higher Likelihood and Lower Prediction Error

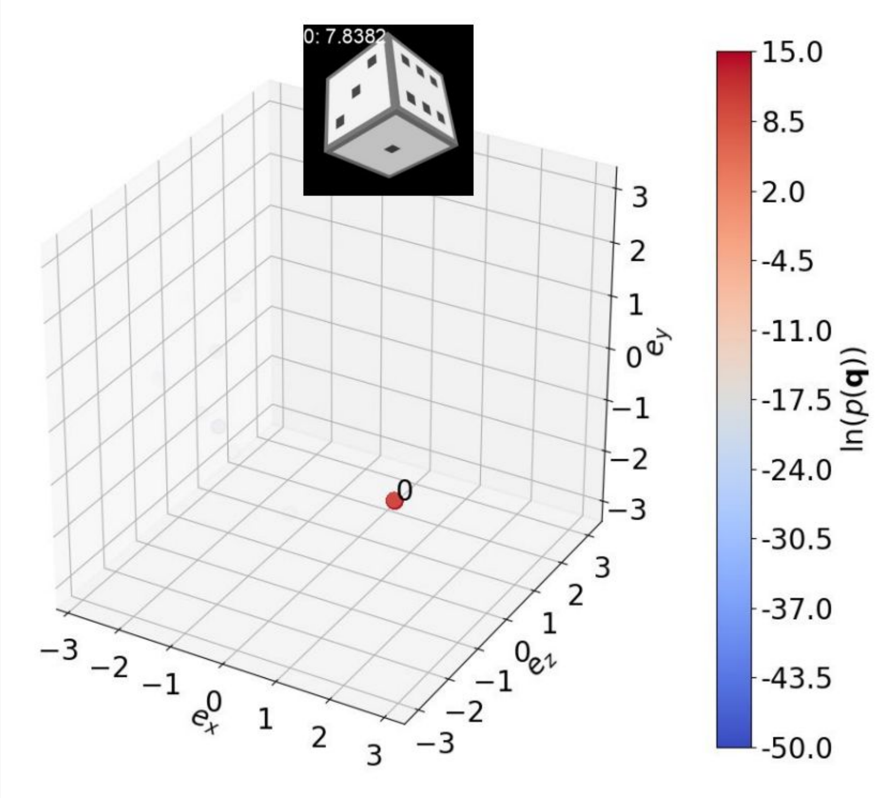
- Trained AQuaMaM from scratch on a large-scale dataset of rendered die images with varying levels of ambiguity.
- Requires generalization
  - Only 135 of the 10,000 test set “quaternion sentences” were seen during training.

Model	**Average LL (↑)**	**Average Distance (↓)**
IPDF	12.29	4.57°
AQuaMaM	**14.01**	**4.32°**

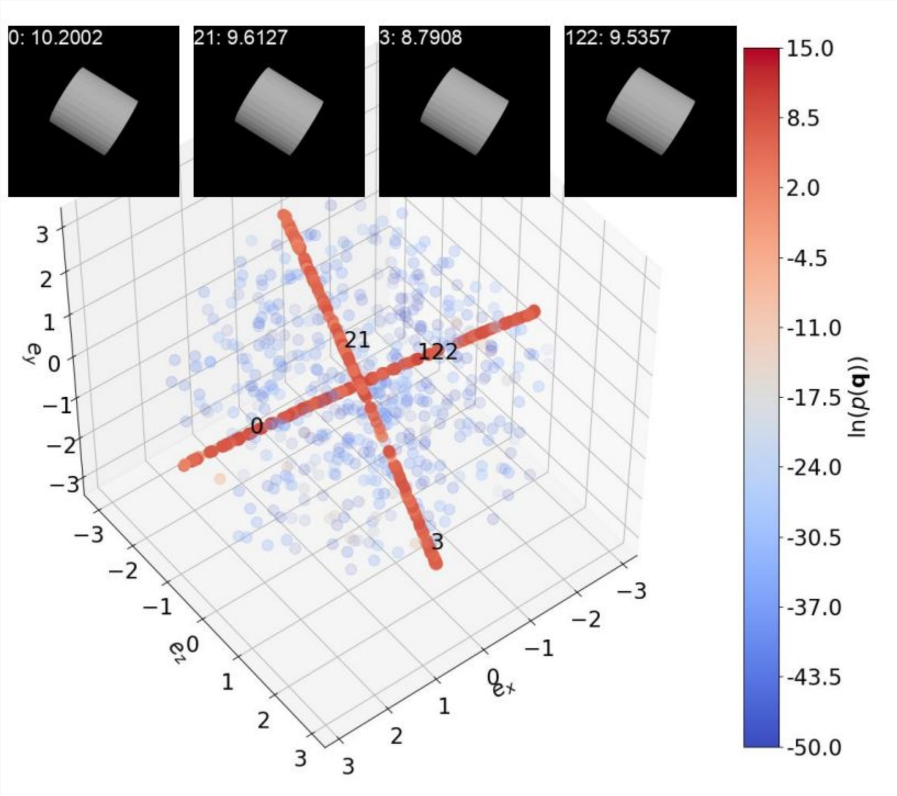
AQuaMaM Precisely Captures Complex, Multimodal Uncertainty



For Unambiguous Views, the Model Correctly Concentrates All Probability at the True Pose



The Framework Extends Naturally to Objects with Continuous Symmetries

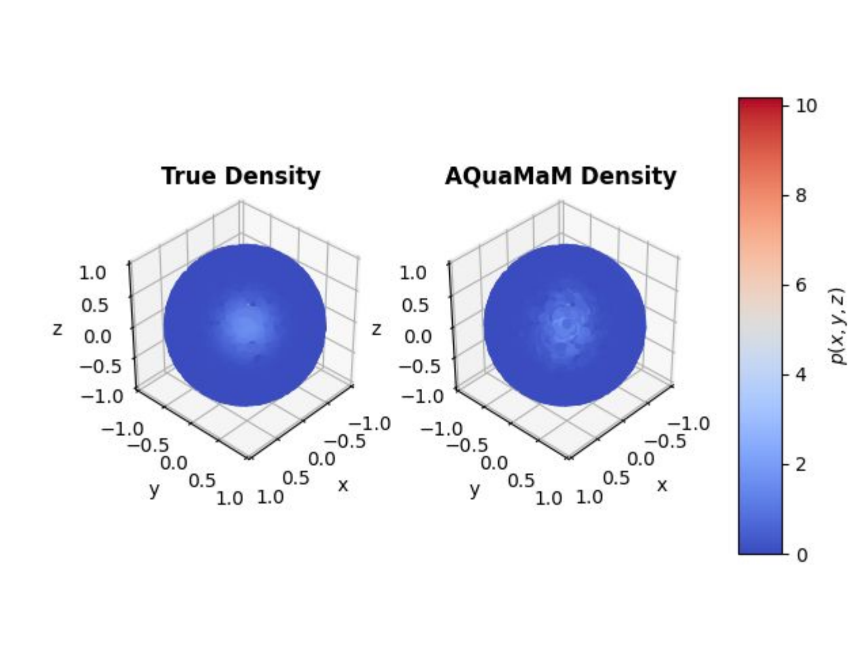


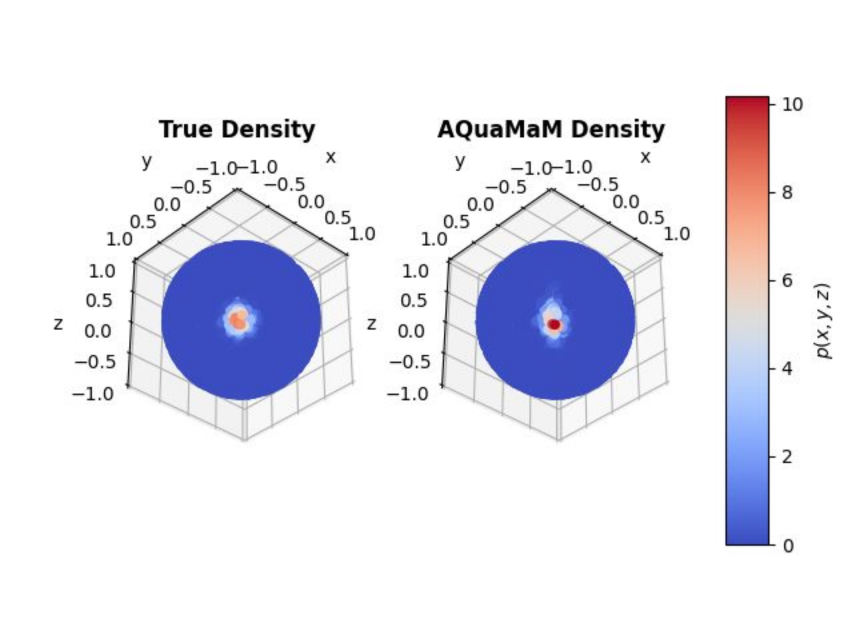
Model	Average LL (↑)
IPDF	5.94
AQuaMaM	**7.24**

And Peak Distributions…

Model	Average LL (↑)
Lieu et al. (2023)	13.93
AQuaMaM	**29.51**

And Spheres…





True density: mixture of two [von Mises-Fisher distributions]

Questions?