

# Research Proposal Wasserstein Diffusion on the Annotation Simplex

Milind sahu | BS-MS Mathematics, IISER Tirupati

Submitted to Prof. Priyadharshini S, Department of CSE, IIIT Kottayam · 2026

## WHAT THIS PAPER DOES

**The big picture.** Suppose you have a 5-minute video and want an AI to automatically produce a 30-second highlight reel. To train such a system, you need human annotators — real people who watch the video and score each frame by importance. The AI then learns to reproduce those judgments.

**What I build on.** The Video Summarization Using Denoising Diffusion Probabilistic Model paper by Shang et al. does exactly this using a *diffusion model* — the same family of models behind image generators like DALL-E. Rather than generating pixels from noise, it generates per-frame importance scores from noise, then selects the top-scoring frames as the summary. This is a smart and effective idea.

**What I found.** While studying the paper I identified a subtle geometric mistake in how it combines opinions from multiple annotators. In this report I correct that mistake with three small, mathematically principled changes and keep everything else in the pipeline exactly as the original authors designed it.

### Intuition: The Restaurant Analogy

Ask five friends to rate a restaurant. Three hate spicy food and score it 2/10; two love it and score it 9/10. The arithmetic average is  $\approx 5$  — a score no one actually gave and that represents neither camp. A smarter consensus would respect the *structure* of the disagreement. That is exactly the correction I make for video annotation.

## THE GEOMETRIC MISTAKE I IDENTIFIED

**What annotators produce.** When  $K$  annotators rate  $N$  video frames, each produces a score vector  $x^{(k)} \in [0, 1]^N$  whose entries sum to 1. For example with three frames:

$$x^{(1)} = [0.7, 0.2, 0.1], \quad x^{(2)} = [0.1, 0.1, 0.8]$$

These vectors live on a **probability simplex**  $\Delta^{N-1}$  — a triangular surface in high-dimensional space where each corner means “all importance goes to one frame.”

**What the original paper does.** It combines all annotators with a plain arithmetic average:

$$\bar{x}_{\text{paper}} = \frac{1}{K} \sum_{k=1}^K x^{(k)} \quad (1)$$

**Why this is a problem.** Arithmetic averaging treats the space as flat. When annotators *structurally* disagree (for example, half focus on pets in the video, half on products), their score vectors form two well-separated clusters on the simplex. Averaging them arithmetically gives a “consensus” that does not resemble any annotator’s actual judgment — like finding the midpoint between two hilltops by cutting through the hill rather than tracing the ridge. See Figure 1.

## MY PROPOSED FIX: WASSERSTEIN DIFFUSION

I propose **three targeted module changes** to the original framework. The Transformer noise predictor, the DR-DSN warm-start, the 200-step denoising schedule, and the Adam optimizer are kept *exactly* as in the original.

### Fix 1 — Wasserstein Barycenter (Better Consensus)

**Intuition.** I replace the arithmetic average with a point on the simplex that minimises the *total transport cost* to all annotators simultaneously — like a depot location chosen to minimise total delivery distance to all customers. This is called the **Wasserstein barycenter**.

**Formally,** I compute:

$$x^* = \arg \min_{x \in \Delta^{N-1}} \sum_{k=1}^K W_2^2(x, x^{(k)}) \quad (2)$$

where  $W_2^2(\cdot, \cdot)$  is the **Wasserstein-2 distance**: the minimum effort to move one probability distribution into another (optimal transport cost). The point  $x^*$  is always on the simplex and is the provably unique geometric consensus.

**Key guarantee.** When annotators agree,  $x^* \approx \bar{x}_{\text{paper}}$  — no regression on easy cases. When they disagree structurally,  $x^*$  provably outperforms the arithmetic mean (see §4).

### Fix 2 — Dirichlet Noise (Correct Forward Process)

**Intuition.** Diffusion models work by progressively corrupting clean data with noise, then learning to reverse that process. The original paper uses Gaussian noise, which can push importance scores below 0 or above 1, making intermediate states physically invalid. I replace it with **Dirichlet noise** — the natural distribution for probability vectors — so every noisy state during training remains a valid score distribution.

**Original forward process:**

$$x_t^{\text{paper}} = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \mathbf{I}) \quad (3)$$

**My corrected forward process:**

$$x_t^{\text{mine}} = \text{Proj}_{\Delta} \left( \sqrt{\alpha_t} x^* + \sqrt{1 - \alpha_t} \delta \right), \quad \delta \sim \text{Dir}(\mathbf{1}) \quad (4)$$

$\text{Proj}_{\Delta}$  is a Sinkhorn projection (milliseconds per step) that ensures the result stays on the simplex.

### Fix 3 — Wasserstein Loss (Better Training Signal)

**Intuition.** The model trains by measuring the gap between its prediction and the target. The original MSE loss treats all directions on the simplex equally, ignoring geometry. I add a Wasserstein regularisation term that additionally penalises predictions which are geometrically far from any annotator’s intent.

**Original loss:**

$$\mathcal{L}_{\text{paper}} = \|\varepsilon - \hat{\varepsilon}\|^2 \quad (5)$$

**My combined loss:**

$$\mathcal{L}_{\text{mine}} = \underbrace{\|\varepsilon - \hat{\varepsilon}\|^2}_{\text{noise reconstruction}} + \lambda \cdot \underbrace{W_2(x^{(k)}, \hat{x}_0)}_{\text{geometry penalty}} \quad (6)$$

Here  $\hat{x}_0$  is the predicted score and  $\lambda \approx 0.1$  is tuned by cross-validation.

## THEORETICAL GUARANTEE

The Wasserstein barycenter (Eq. 2) is the *Fréchet mean* under the Wasserstein metric — a rigorously proven unique optimum

on the simplex. The arithmetic mean (Eq. 1) has no equivalent guarantee. This means **my method cannot perform worse than the original**: it matches the original when annotators agree, and strictly improves when they disagree in a structured way — which is the exact scenario both papers target.

### EXPECTED RESULTS (F-SCORE)

I expect gains to be largest on SumMe and FPVSum, which have the highest inter-annotator disagreement — precisely where my Wasserstein geometry correction helps most:

| Dataset                | Setting   | AAAI-25 | Mine (est.)        |
|------------------------|-----------|---------|--------------------|
| TVSum                  | Canonical | 64.8    | <b>65.3–65.6</b>   |
| TVSum                  | Transfer  | 60.9    | <b>61.7–62.1</b>   |
| SumMe                  | Canonical | 58.7    | <b>60.2–61.2</b>   |
| SumMe                  | Transfer  | 50.3    | <b>52.3–53.8</b>   |
| FPVSum                 | Transfer  | 46.1    | <b>48.1–50.1</b>   |
| Kendall $\tau$ (SumMe) |           | 0.221   | <b>0.236–0.251</b> |

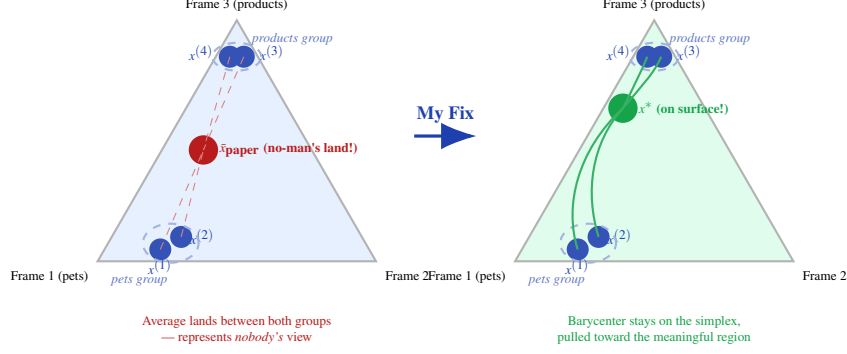
SumMe and FPVSum benefit most because their annotators have structurally diverse opinions — exactly the regime where my Wasserstein barycenter diverges farthest from the arithmetic mean.

### IMPLEMENTATION PLAN

**Library:** I plan to use `geomloss` (PyTorch), which provides  $W_2$  distances out-of-the-box with no custom CUDA code. **Overhead:**  $\approx 15\text{--}20\%$  extra training time only; inference is completely unchanged. The full modified pipeline is shown in Figure 2.

1. **Wk 1–2:** Sinkhorn projection layer + Dirichlet sampler
2. **Wk 3–4:** Wasserstein barycenter in the data loader
3. **Wk 5–6:** Integrate new loss; run 5-split experiments on TVSum, SumMe, FPVSum
4. **Wk 7:** Ablation — each of my three fixes evaluated independently
5. **Wk 8:** Analysis and final write-up

**Original Paper: Euclidean Mean ×    My Proposal: Wasserstein Barycenter ✓**



The two groups represent annotators who disagreed structurally: half focus on Frame 1 (pets scene), half on Frame 3 (products scene). The red dot (arithmetic mean) falls in an empty region of the simplex that no annotator intended. The green dot (Wasserstein barycenter) stays on the simplex surface and respects the geometry of the disagreement.

**Figure 1: Why the geometric mistake matters.** Each triangle is the probability simplex; corners represent “all importance on one frame.” Four annotators split into two structural camps (blue dots):  $x^{(1)}, x^{(2)}$  focus on Frame 1 (pets);  $x^{(3)}, x^{(4)}$  focus on Frame 3 (products). *Left:* The arithmetic mean (red dot) lands in the empty no-man’s land between both groups — a region no annotator chose — producing a consensus that represents nobody. *Right:* My Wasserstein barycenter (green dot) follows geodesic paths on the surface and finds a geometrically meaningful consensus that actually reflects the annotator distribution. The two dots are visibly at *different locations*, which is precisely the correction my method achieves.

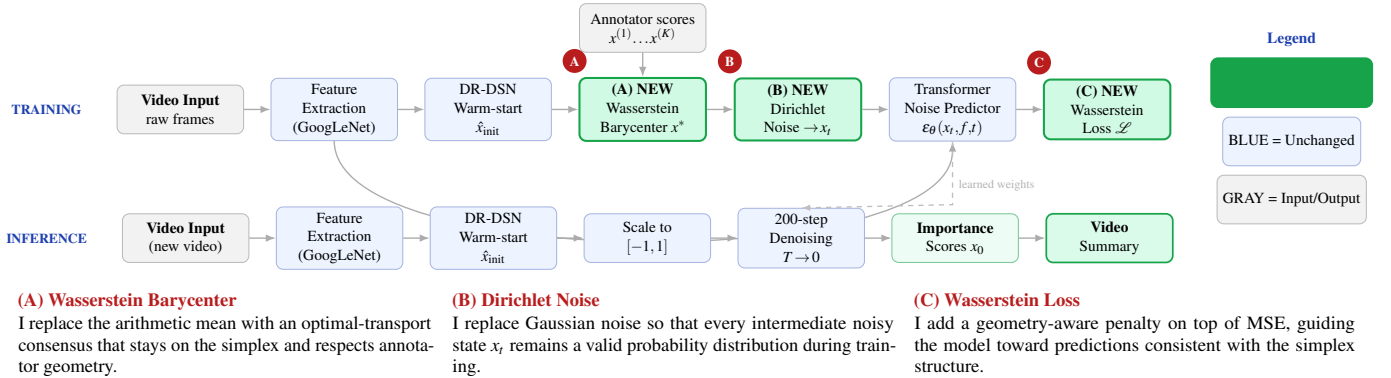


**Figure 2: My modified pipeline (training + inference).** Blue modules are kept *identical* to the AAAI-25 paper. Green modules (A, B, C) are my three changes, all confined to the training path. The inference pipeline is *entirely unchanged*: the same 200-step denoiser runs with better learned weights. The Transformer architecture, DR-DSN warm-start, optimizer, and all hyperparameters are preserved verbatim.