CrossFlowDG: Bridging the Modality Gap with Cross-modal Flow Matching for Domain Generalization

Supplementary Material

A. Textual Domain Bank Entries

Table S1 enumerates the $k = 18$ prompt templates used to construct the Textual Domain Bank (TDB), as described in Section 6.1.

Table S1. Templates used in the Textual Domain Bank. Each template is completed with the target class name.

#	Template
1	a picture of a [class]
2	an image of a [class]
3	a photograph of a [class]
4	a painting of a [class]
5	a sketch of a [class]
6	a cartoon of a [class]
7	a 3D render of a [class]
8	a drawing of a [class]
9	a grayscale image of a [class]
10	a low-light image of a [class]
11	a high-resolution image of a [class]
12	a blurred image of a [class]
13	an overexposed image of a [class]
14	a noisy image of a [class]
15	a close-up image of a [class]
16	a wide-angle image of a [class]
17	an indoor image of a [class]
18	an outdoor image of a [class]

B. Ablation on ODE Integration Steps

We investigate the trade-off between classification accuracy and inference efficiency by evaluating *CrossFlowDG* across varying numbers of ODE integration steps, $N \in \{1, 6, 12\}$. The results are summarized in Table S2.

Table S2. Ablation on the number of ODE integration steps (N). **Bold** indicates best.

N	L100	L38	L43	L46	Avg.
1	66.5	49.1	62.4	48.5	56.6
6	**67.2**	48.1	62.5	47.9	56.4
12	66.6	**52.0**	**62.8**	**50.4**	**58.0**

Notably, the single-step configuration ($N = 1$) achieves an average accuracy of 56.6%, namely a 1.4% drop compared to the full 12-step model. It performs particularly well on the quantitatively easier domains (L100, L43), achieving accuracies comparable to the 12-step baseline. The 1-step performance empirically suggests that the learned vector field is highly consistent with the linear interpolation objective of flow matching, meaning a single Euler step provides a strong approximation of the true transport trajectory. While $N = 6$ yields marginal improvements on the easier domains, it suffers a performance drop on the more challenging ones (L38, L46). The full $N = 12$ configuration yields higher accuracy in three out of four domains, and the highest overall accuracy.

C. Inference Efficiency

Table S3 confirms that the lightweight flow model contributes minimally to the overall computational footprint; the dominant cost remains the VMamba-T backbone, which is executed exactly once per sample. Scaling the integration steps from $N = 1$ to $N = 12$ adds approximately 3.6 ms of overhead (11.47 ms vs. 15.13 ms on a consumer-grade NVIDIA RTX 4050 GPU). Because the flow map parameters (~1M) remain easily cached in VRAM, the iterative memory bottlenecks are largely bypassed. These baseline latency measurements on a consumer GPU indicate viability for resource-constrained edge deployment. Future work could explore flow distillation [7, 8] to achieve multi-step accuracy at the computational cost of a single-step forward pass.

Table S3. Computational cost and inference latency across $N \in \{1, 6, 12\}$ ODE integration steps. Latency is measured as the average over 10,000 forward passes of a single sample on an NVIDIA RTX 4050 GPU.

N	GFLOPs	Latency (ms)
1	4.92	11.47
6	4.98	13.22
12	5.06	15.13