TIGHTENING MULTIMODAL REGRESSION GENERALIZATION BOUNDS VIA STRUCTURED PREDICTIVE DISTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal regression is increasingly being deployed in real-world applications, yet its generalization performance remains limited, particularly under noisy conditions. Existing methods predominantly rely on point-wise predictions, leaving the role of structured predictive representations in generalization-bound tightening underexplored. In this work, we propose **S**tructured **P**redictive **D**istributions (**SPD**) for tightening generalization bounds in multimodal regression, a theoretically grounded framework that reformulates multimodal regression in a structured predictive distribution space. Concretely, SPD combines Normal-Inverse-Gamma predictive representations with regularization to structure continuous regression outputs and performs multimodal fusion within this structured space. We further introduce a distribution-aware reweighting mechanism that adapts optimization to the predictive representation to improve robustness under noisy conditions. Using a common squared error risk, our analysis establishes tighter generalization bounds through NIG-induced capacity control. We further characterize the empirical-risk condition for comparisons with a matched reference predictor, relating empirical fit to the reduction in the complexity term. Extensive experiments on multiple multimodal regression benchmarks, covering image fusion, crowd counting, and depth estimation, demonstrate that SPD consistently improves robustness under diverse noise conditions. Our code is available in the supplementary material and `https://anonymous.4open.science/r/SPD-A356/`.

1 INTRODUCTION

Multimodal regression has garnered substantial significance in recent years, particularly in tasks such as depth estimation Ming et al. (2021); Li & Snavely (2018); Zhu et al. (2019), crowd counting Lin et al. (2021; 2022; 2024), and sensor fusion Wang et al. (2019). However, multimodal models are often exposed to low-quality data with diverse noisy signals, leading to degraded robustness. To address this issue, Robust Multimodal Learning (RML) has been extensively studied. While existing RML methods Wei et al. (2025); Hu et al. (2026); Zhang et al. (2023) achieve promising robustness in classification tasks due to their structured predictive space, e.g., simplex-constrained outputs, such strategies are not directly applicable to regression settings, where predictions lie in an unconstrained continuous space without intrinsic structural regularization.

Several works have attempted to improve the performance of multimodal regression by enhancing feature-level fusion Zhang et al. (2025); Meng et al. (2025; 2024); Devulapally et al. (2024). However, as illustrated in Fig. 1(a), these methods focus on point-wise prediction, while the role of structured predictive representations in hypothesis-class control remains less explored. Point-wise supervision fits target values but does not by itself impose a distributional structure on the predictive representation. This motivates studying whether explicit predictive regularization can control the hypothesis class and tighten generalization bounds. As a result, these approaches often exhibit limited robustness and suboptimal generalization performance under challenging multimodal conditions, as empirically illustrated in Fig. 1(b). More importantly, existing multimodal regression methods lack a principled theoretical understanding of how predictive space affects generalization. We address

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.

this gap by reformulating multimodal regression in a structured predictive space and providing a theoretical framework that explains its generalization properties.

To this end, we propose **S**tructured **P**redictive **D**istributions (**SPD**), a unified framework that reformulates multimodal regression in a structured predictive space with intrinsic geometric constraints. Specifically, SPD combines a Normal-Inverse-Gamma parameterization with predictive regularization to structure continuous outputs. Building on established NIG-based evidential regression, our contribution lies in integrating structure preserving multimodal fusion, predictive regularization, and distribution-aware reweighting within a unified framework. We further develop a Rademacher complexity analysis linking NIG regularization to capacity control of the final point-prediction class and conditional generalization-bound tightening. We show that, by imposing structural constraints, SPD reduces the effective complexity of the hypothesis class and leads to tighter generalization bounds. In addition, we introduce a distribution-aware reweighting mechanism (DAR) that dynamically adjusts the learning process according to the predictive structure, enabling more robust optimization under challenging multimodal conditions and further contributing to improved generalization behavior. Our contributions can be summarized as follows:

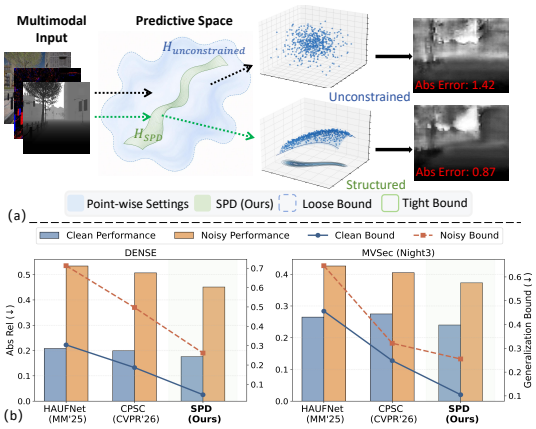

Figure 1: (a) Illustration of point-wise and structured SPD predictions, with smaller errors for SPD in the depicted example, motivating the analysis of generalization-bound tightening. (b) Depth estimation performance of point-wise baselines and SPD, together with the corresponding empirical bound proxies.

- We reformulate multimodal regression in a structured predictive space, enabling structure-preserving multimodal fusion and explicit regularization of predictive representations.
- We propose distribution-aware reweighting (DAR) to adaptively modulate learning, improving robustness and optimization stability.
- We derive Rademacher bounds for a common squared error risk, establishing generalization-bound tightening and an empirical-risk criterion for learned predictors.

2 Related Work

Robust Multimodal Learning. Multimodal models integrate heterogeneous modalities to learn complementary representations, thereby achieving substantial performance improvements over unimodal approaches. Yet, practical multimodal learning must contend with imperfect data and real-world conditions, necessitating enhanced robustness and tighter generalization boundaries. Previous works made a lof of efforts to enhance the robustness: Zhang et al.Zhang et al. (2023) and Cao et al.Cao et al. (2024a) adapted dynamic fusion to improve the generalization boundary, gaining better robustness; Yang et al.Yang et al. (2024) proposed a certifiable training strategy to enhance robustness against modality-specific perturbations; Gao et al.Gao et al. (2024b), Hu et al.Hu et al. (2026), and Gao et al.Gao et al. (2024a) explicitly modeled the uncertainty from the Gaussian in multimodal models and leverage it as the calibrations of the modal representations. Despite their effectiveness, most of them are designed for classification and rely on fixed predictive spaces such as logits, probabilistic simplex or Gaussian forms, which limits their ability to characterize continuous regression outputs. In contrast, we reformulate multimodal regression in a structured predictive space, which are used as constrained output representations to regulate the hypothesis space and improve generalization against the modal perturbations.

Predictive Space Construction. The design of predictive space plays an important role in shaping model behavior and generalization. Recent probabilistic regression methods, such as Deep Evidential Regression Amini et al. (2020), introduce Normal-Inverse-Gamma parameterizations

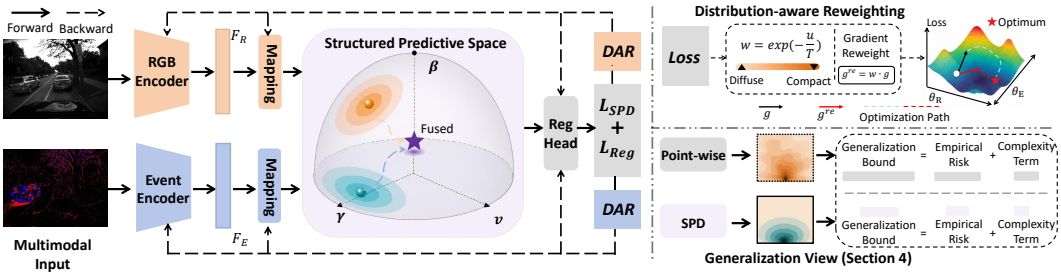

Figure 2: SPD maps multimodal features into a structured predictive space, and performs structure-preserving integration to obtain a fused predictive representation. Distribution-aware reweighting (DAR) adaptively modulates optimization, while the generalization view illustrates that SPD reduces effective complexity and tightens the generalization bound compared with point-wise regression.

to enrich regression outputs, and subsequent works further improve their stability through regularization strategies Ye et al. (2024); Wu et al. (2024); Oh & Shin (2022); Meinert et al. (2023); Ma et al. (2021). However, these methods mainly focus on predictive uncertainty estimation and calibration. Different from them, our method treats predictive distributions as structured output representations for constructing a constrained predictive space. Within this space, we integrate structure-preserving multimodal fusion, predictive regularization, and distribution-aware reweighting to improve robustness under noisy inputs. We further analyze how NIG regularization controls the point-prediction class under explicit assumptions, connecting the structured formulation to conditional generalization-bound tightening.

3 Method

3.1 Problem Setup

Suppose there are two heterogeneous modalities, denoted as $\phi_1 \in \mathbb{R}^{d_1}$ and $\phi_2 \in \mathbb{R}^{d_2}$, with ground-truth labels $(y_1, y_2, \ldots, y_N) \in \mathcal{Y} \subseteq \mathbb{R}^d$. Multimodal regression models aim to learn a mapping from multimodal inputs to the target space by first extracting modality-specific features $\mathcal{F}_m = \mathcal{E}_m(\phi_m),\ m \in \{1, 2\}$, and then producing the final prediction via a regression head: $\hat{y} = \mathcal{O}([\mathcal{F}_1; \mathcal{F}_2])$, where $\mathcal{E}_m(\cdot)$ denotes modality-specific encoder and $\mathcal{O}(\cdot)$ denotes a learnable regression predictor. The model is typically optimized by minimizing a point-wise regression loss: $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \ell(\hat{y}^{(i)}, y^{(i)})$, with gradients given by $\nabla_\theta \mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \nabla_\theta \ell(\hat{y}^{(i)}, y^{(i)})$. As illustrated in the introduction, this formulation implicitly assumes an unconstrained predictive space, leading to a less controlled hypothesis space and suboptimal generalization behavior.

3.2 Structured Predictive Distributions

We introduce a structured predictive mapping that first maps multimodal inputs into constrained predictive representations and then derives the final regression output after structure-preserving integration. Compared with directly mapping features to unstructured scalar predictions, this formulation imposes intrinsic constraints before prediction. Formally, instead of directly learning a point-wise predictor $f : \mathcal{X} \to \mathbb{R}$, we define:

$$f_{\mathcal{P}} : \mathcal{X} \to \mathcal{P}, \tag{1}$$

where $\mathcal{X}$ denotes the multimodal input space and $\mathcal{P}$ denotes the space of valid predictive representations endowed with intrinsic constraints.

To realize the structured predictive space $\mathcal{P}$, we adopt a parametric representation. In this work, we employ a Normal-Inverse-Gamma (NIG) parameterization as an instantiation of the structured predictive representation. Specifically, for the i-th sample and modality m, given the modality-specific feature representation $\mathcal{F}_m^{(i)}$, we map it into the structured predictive space via a learnable transformation $g_m(\cdot)$:

$$\mathbf{z}_m^{(i)} = (\gamma_m^{(i)}, \upsilon_m^{(i)}, \alpha_m^{(i)}, \beta_m^{(i)}) = g_m(\mathcal{F}_m^{(i)}) = g_m(\mathcal{E}_m(\phi_m^{(i)})). \tag{2}$$

Here, $\mathbf{z}_m^{(i)}$ denotes the modality-specific structured predictive representation. To ensure valid representations within the structured predictive space, we enforce:

$$\upsilon_m^{(i)} > 0, \quad \alpha_m^{(i)} > 1, \quad \beta_m^{(i)} > 0, \tag{3}$$

which are implemented via appropriate activation functions, such as softplus. Given the modality-specific structured predictive representations $\{\mathbf{z}_m^{(i)}\}_{m=1}^M$, we perform multimodal integration within the structured predictive space:

$$\mathbf{z}_f^{(i)} = \mathcal{I}_{\mathcal{P}}\left(\{\mathbf{z}_m^{(i)}\}_{m=1}^M\right), \quad \mathbf{z}_f^{(i)} = (\gamma_f^{(i)}, \upsilon_f^{(i)}, \alpha_f^{(i)}, \beta_f^{(i)}), \tag{4}$$

where $\mathcal{I}_{\mathcal{P}}(\cdot)$ denotes a structure preserving integration operator that maps modality-specific representations into a valid fused representation in $\mathcal{P}$. The detail of $\mathcal{I}_{\mathcal{P}}(\cdot)$ is provided in the Appendix.

Since multimodal regression requires a continuous prediction, we obtain the final output by applying a regression readout to the location parameter of the fused structured representation:

$$\hat{y}^{(i)} = \mathcal{O}(\gamma_f^{(i)}). \tag{5}$$

For likelihood-based supervision, we use $\hat{y}^{(i)}$ as the location parameter of the target-space predictive distribution, together with the corresponding fused parameters. $\upsilon_f^{(i)}$, $\alpha_f^{(i)}$, and $\beta_f^{(i)}$. These parameters are aligned with the target coordinates and satisfy the NIG validity conditions.

In this setting, the supervision loss is the NIG-induced predictive negative log-likelihood:

$$\mathcal{L}_{\mathrm{SPD}}^f = -\frac{1}{N}\sum_{i=1}^{N} \log p\left(y^{(i)} \mid \hat{y}^{(i)}, \upsilon_f^{(i)}, \alpha_f^{(i)}, \beta_f^{(i)}\right). \tag{6}$$

In addition to the fused predictive likelihood, we regularize the modality-specific structured representations by encouraging them to remain close to a fixed reference representation:

$$\mathcal{L}_{\mathrm{reg}} = \frac{1}{M}\sum_{m=1}^{M} \mathrm{D}\big((\gamma_m, \upsilon_m, \alpha_m, \beta_m), (\gamma_0, \upsilon_0, \alpha_0, \beta_0)\big). \tag{7}$$

Here, $(\gamma_0, \upsilon_0, \alpha_0, \beta_0)$ denotes the reference representation, and $\mathrm{D}(\cdot,\cdot)$ measures the discrepancy between structured predictive representations. The averaging over training samples is implicit.

Combining the task-specific loss with this regularizer gives the training objective:

$$\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \lambda \mathcal{L}_{\mathrm{reg}}, \tag{8}$$

where λ controls the regularization strength. For likelihood-based supervision, $\mathcal{L}_{\mathrm{task}} = \mathcal{L}_{\mathrm{SPD}}^f$ as defined in Eq. (6). For image fusion, we use the source-image-based fusion objective described in Appendix A.2. The framework is illustrated in Fig. 2.

3.3 Distribution-Aware Reweighting

In addition to the structured predictive formulation, we further introduce a distribution aware modulation (DAR) to adaptively adjust the learning process. Specifically, we acquire a modulation signal from the structured predictive representation, which characterizes the local spread of the predictive space geometry:

$$\begin{aligned} s_m^{(i)} &= \mathrm{mean}\left[\frac{\beta_m^{(i)}}{\alpha_m^{(i)} - 1} + \frac{\beta_m^{(i)}}{\upsilon_m^{(i)}(\alpha_m^{(i)} - 1)}\right], \\ w_m^{(i)} &= \exp\left(-\frac{s_m^{(i)}}{T}\right), \qquad T > 0. \end{aligned} \tag{9}$$

where T is a temperature parameter. This design assigns smaller weights to representations with larger geometric spread. The corresponding modulated gradient with respect to the model parameters θ_m is given by:

$$\bar{w}_m = \frac{1}{n}\sum_{i=1}^{n} \mathrm{stopgrad}(w_m^{(i)}), \qquad g_m^{\mathrm{DAR}} = \bar{w}_m \nabla_{\theta_m} \mathcal{L}_{\mathcal{B}}. \tag{10}$$

Here, n is the mini-batch size, $\mathcal{L}_{\mathcal{B}}$ is the total training objective, and θ_m denotes the modality-specific NIG head weights. DAR uses predictive spread as a learned uncertainty signal to adapt the model optimization.

4 Theoretical Analysis

For the theoretical analysis, we consider regression with $\mathcal{S} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N} \sim \mathcal{D}^N$, where $y \in \mathbb{R}^d$ and $h(x)$ denotes the final point prediction. We compare generalization using a common squared error risk. The NIG objective and DAR govern optimization, while this common evaluation loss makes the resulting predictors directly comparable.

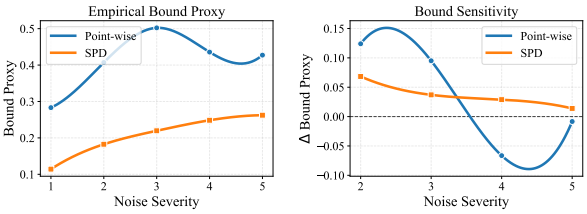

Figure 3: Empirical bound-proxy diagnostics on DENSE under varying noise severity. Left: proxy values for point-wise regression and SPD. Right: changes in the proxy between adjacent noise-severity levels.

Concretely, We analyze the generalization bound by examining its empirical-risk and hypothesis-class complexity terms. Specifically, we study the conditions under which NIG-induced capacity control yields tighter bounds, accounting for empirical-risk differences between learned predictors.

Definition 1. (***Common Regression Risk***) *For $\ell(h, x, y) = \|h(x) - y\|_2^2$, we define*:

$$\mathcal{R}_\ell(h) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(h, X, Y)],$$
$$\widehat{\mathcal{R}}_\ell(h) = \frac{1}{N}\sum_{i=1}^{N} \ell(h, x^{(i)}, y^{(i)}). \tag{11}$$

Let $\mathcal{H}$ denote a class of point predictors $h : \mathcal{X} \to \mathbb{R}^d$. To analyze generalization, we introduce the empirical Rademacher complexity of the associated loss class.

Definition 2. (***Rademacher Complexity***) *For the loss ℓ defined in Definition 1, the empirical Rademacher complexity of $\ell \circ \mathcal{H}$ is:*

$$\widehat{\mathfrak{R}}_{\mathcal{S}}(\ell \circ \mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N} \sigma_i \ell(h, x^{(i)}, y^{(i)})\right], \tag{12}$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)$ are independent Rademacher variables satisfying $\Pr(\sigma_i = +1) = \Pr(\sigma_i = -1) = 1/2$.

In the following, we analyze the generalization behavior of the proposed framework by introducing the following lemma, which provides a bound on the expected risk in terms of the empirical risk and the Rademacher complexity.

Lemma 1. (***Generalization Bound***) *Under the bounded-domain conditions in Appendix D, assume that $\|Y\|_2 \le B_y$ almost surely and $\sup_{x\in\mathcal{X}} \|h(x)\|_2 \le B_h$ for every predictor in either comparison class, where B_y and B_h are fixed independently of the sample $\mathcal{S}$. For the common squared-error loss in Definition 1, set $B_\ell = (B_h + B_y)^2$ and $L = 2(B_h + B_y)$. Then $0 \le \ell(h, x, y) \le B_\ell$ almost surely, and each scalar coordinate loss is L-Lipschitz on the stated domain. For vector-valued predictions, define $\mathcal{H}_j = \{x \mapsto h_j(x) : h \in \mathcal{H}\}$ and*

$$\mathfrak{C}_{\mathcal{S}}(\mathcal{H}) = \sum_{j=1}^{d} \widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{H}_j).$$

For any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over $\mathcal{S} \sim \mathcal{D}^N$, the following holds simultaneously for both sample-independent comparison classes and all predictors h in each class:

$$\begin{aligned}\mathcal{R}_\ell(h) &\le \widehat{\mathcal{R}}_\ell(h) + 2\widehat{\mathfrak{R}}_{\mathcal{S}}(\ell \circ \mathcal{H}) + C_{N,\delta} \\ &\le \widehat{\mathcal{R}}_\ell(h) + 2L\mathfrak{C}_{\mathcal{S}}(\mathcal{H}) + C_{N,\delta},\end{aligned} \tag{13}$$

Table 1: Comparison under clean and noise settings for Image Fusion. Results are averaged over LLVIP, MEF, and MFF datasets.

Method	Clean				Noise			
	SSIM↑	MS↑	QW↑	FMI↑	SSIM↑	MS↑	QW↑	FMI↑
CCF (NIPS'24)	0.501	0.941	0.732	0.880	0.338	0.814	0.786	0.856
EAU (CVPR'24)	0.516	0.911	0.835	0.862	0.338	0.806	0.774	0.841
PDF (ICML'24)	0.509	0.931	0.858	0.864	0.340	0.806	0.782	0.837
CRMT (ICLR'24)	0.523	0.946	0.869	0.879	0.347	0.819	0.778	0.858
TITA (ICCV'25)	0.537	0.963	0.878	**0.894**	0.342	0.828	0.796	0.867
HOVQ (AAAI'26)	0.452	0.828	0.718	0.814	0.293	0.741	0.749	0.803
CPSC (CVPR'26)	0.502	0.947	0.843	0.835	0.313	0.785	0.774	0.824
SPD (Ours)	**0.542**	**0.964**	**0.882**	**0.894**	**0.406**	**0.836**	**0.802**	**0.869**

Table 2: Quantitative comparison on the DENSE dataset under clean and noise settings.

Method	Clean						Noise					
	10m↓	20m↓	30m↓	Abs.↓	R_{log} ↓	SI_{log} ↓	10m↓	20m↓	30m↓	Abs.↓	R_{log} ↓	SI_{log} ↓
ER-F2D (CVPR'24)	0.728	2.120	2.845	0.246	0.395	0.082	1.623	3.772	4.638	0.622	0.824	0.559
PDF (ICML'24)	0.696	1.919	2.621	0.215	0.371	0.081	1.468	3.540	4.453	0.495	0.836	0.534
CRMT (ICLR'24)	1.005	1.863	2.424	0.226	0.308	0.078	1.592	3.520	4.403	0.492	0.765	0.496
EAU (CVPR'24)	0.697	1.846	2.568	0.228	0.359	0.091	1.759	3.542	4.588	0.512	0.862	0.522
HAFUNet (MM'25)	0.650	1.944	2.655	0.208	0.381	0.089	1.727	3.481	4.332	0.534	0.763	0.493
HOVQ (AAAI'26)	0.608	1.675	2.279	0.179	0.308	0.073	1.639	3.687	4.569	0.564	0.844	0.480
CPSC (CVPR'26)	0.682	1.741	2.357	0.199	0.326	0.075	1.561	3.482	4.366	0.507	0.826	0.507
SPD (Ours)	**0.581**	**1.582**	**2.203**	**0.176**	**0.299**	**0.070**	**1.450**	**3.385**	**4.266**	**0.451**	**0.745**	**0.465**

where $C_{N,\delta} = 3B_\ell\sqrt{\log(4/\delta)/(2N)}$. The second inequality follows scalar contraction applied coordinatewise. For $d = 1$, $\mathfrak{C}_S(\mathcal{H})$ reduces to the usual predictor-class Rademacher complexity.

With a common loss range and confidence term, Lemma 1 separates the generalization bound into an empirical-risk term and a complexity term. We next establish conditional generalization-bound tightening through NIG-induced capacity control and characterize the empirical-risk condition for comparison with a matched reference predictor.

Fig. 3 reports an empirical bound proxy and its signed variation across noise levels. The proxy combines the common empirical squared error risk with a nonnegative residual-based term. Under the conditions of Corollary 1, combining this proxy with the NIG-induced capacity term yields a conservative generalization upper bound. Its definition, computation, and connection to the theoretical analysis are provided in Appendix C.8.

Corollary 1. **(*Conditional Generalization-Bound Tightening*)** *Under Assumption 1 in the Appendix, let $\mathcal{H}_{\mathrm{SPD}}^{\mathrm{point}}$ denote the constrained SPD point-prediction class, and let $\mathcal{H}_G^{\mathrm{point}}(R)$ denote its matched Gaussian reference class. Suppose $r_{\mathrm{NIG}} \le R$, and define:*

$$\rho = \frac{r_{\mathrm{NIG}}}{R}, \qquad \Gamma_G = 2L\mathfrak{C}_S\big(\mathcal{H}_G^{\mathrm{point}}(R)\big). \tag{14}$$

For any $\delta \in (0,1)$, with probability at least $1-\delta$, every $h \in \mathcal{H}_{\mathrm{SPD}}^{\mathrm{point}}$ satisfies:

$$\begin{aligned}\mathcal{R}_\ell(h) &\le \widehat{\mathcal{R}}_\ell(h) + \rho\Gamma_G + C_{N,\delta} \\ &\le \widehat{\mathcal{R}}_\ell(h) + \Gamma_G + C_{N,\delta},\end{aligned} \tag{15}$$

where $\mathfrak{C}_S$, L, and $C_{N,\delta}$ are defined in Lemma 1.

The two upper bounds evaluate the same predictor using the constrained SPD class and the matched reference class, respectively. The bound reduction is $(1-\rho)\Gamma_G$, which is strictly positive when $\rho < 1$ and $\Gamma_G > 0$. The complete class definitions, the derivation of r_{NIG}, and the proof are provided in the Appendix.

To clarify the bound comparison, we examine the empirical-risk and complexity terms separately. Complete proofs are provided in the Appendix.

Table 3: Quantitative comparison on MVSEC under clean and noise settings. **More results can be found in appendix.**

Method	Clean						Noise					
	night1		night2		night3		night1		night2		night3	
	Abs.↓	R_{log} ↓	Abs.↓	R_{log} ↓	Abs.↓	R_{log} ↓	Abs.↓	R_{log} ↓	Abs.↓	R_{log} ↓	Abs.↓	R_{log} ↓
ER-F2D (CVPR'24)	0.329	**0.421**	0.287	0.465	0.247	0.452	0.490	0.915	0.530	0.802	0.463	0.896
PDF (ICML'24)	0.505	0.619	0.358	0.522	0.336	0.521	0.640	0.877	0.533	0.819	0.572	0.855
CRMT (ICLR'24)	0.596	0.522	0.459	0.422	0.360	0.414	0.641	0.658	0.467	0.582	0.444	0.582
EAU (CVPR'24)	0.425	0.523	0.312	0.488	0.289	0.477	0.541	0.824	0.461	0.798	0.522	0.823
HAFUNet (MM'25)	0.329	0.469	0.282	0.424	0.261	0.429	0.468	0.811	0.426	0.770	0.426	0.781
HOVQ (AAAI'26)	0.442	0.551	0.280	0.439	0.265	0.463	0.568	0.997	0.597	0.908	0.583	0.929
CPSC (CVPR'26)	0.354	0.506	0.275	0.479	0.275	0.479	0.483	0.737	0.409	0.697	0.405	0.716
SPD (Ours)	**0.326**	0.457	**0.262**	**0.403**	**0.240**	**0.405**	**0.398**	**0.633**	**0.402**	**0.562**	**0.373**	**0.563**

(i) Empirical risk term. For the same predictor, the empirical-risk term remains unchanged in the two bounds of Corollary 1. For comparison across different predictors, let $h_{\mathrm{SPD}} \in \mathcal{H}_{\mathrm{SPD}}^{\mathrm{point}}$ and $h_G \in \mathcal{H}_G^{\mathrm{point}}(R)$ denote the learned SPD and reference predictors, respectively. Using the common squared error loss in Definition 1, define their empirical-risk difference as:

$$\Delta_{\mathrm{emp}} = \widehat{\mathcal{R}}_\ell(h_{\mathrm{SPD}}) - \widehat{\mathcal{R}}_\ell(h_G). \tag{16}$$

The difference between the SPD upper bound and the reference upper bound is then $\Delta_{\mathrm{emp}} - (1-\rho)\Gamma_G$, since both bounds use the same confidence term. Consequently, the SPD upper bound is no larger than the reference upper bound whenever $\Delta_{\mathrm{emp}} \le (1-\rho)\Gamma_G$. Thus, generalization-bound tightening across predictors is obtained when any increase in empirical risk is offset by the reduction in the complexity term. This comparison applies to learned predictors without assuming exact empirical-risk minimization. The derivation is provided in the Appendix.

(ii) Complexity term. Under Assumption 1 in the Appendix, the NIG regularization budget, together with the variance and feature conditions, bounds the norm of the location head parameters by r_{NIG}. The matched fusion and readout architecture then yields inclusion in a smaller-radius subclass of the reference class. For $r_{\mathrm{NIG}} \le R$, this gives:

$$2L\mathfrak{C}_{\mathcal{S}}\big(\mathcal{H}_{\mathrm{SPD}}^{\mathrm{point}}\big) \le \rho\Gamma_G \le \Gamma_G, \tag{17}$$

where $\rho = r_{\mathrm{NIG}}/R$ and Γ_G are defined in Corollary 1. Both sides use the common squared error loss and the same Lipschitz constant L. As a result, combining this capacity control with Lemma 1 yields Corollary 1.

(iii) Effect of distribution-aware reweighting. As specified in Eq. 10, As specified in Eq. (10), DAR scales the mini-batch weight gradient of each modality-specific NIG head using an aggregate of temperature controlled coefficients derived from predictive spread. We analyze the resulting predictor under the class conditions of Corollary 1. Samples with larger predictive spread receive smaller weights in the update.

Let h_{DAR} denote the final predictor obtained using DAR. Provided that $h_{\mathrm{DAR}} \in \mathcal{H}_{\mathrm{SPD}}^{\mathrm{point}}$ and the conditions of Corollary 1 hold, the same generalization guarantee applies:

$$\mathcal{R}_\ell(h_{\mathrm{DAR}}) \le \widehat{\mathcal{R}}_\ell(h_{\mathrm{DAR}}) + \rho\Gamma_G + C_{N,\delta}. \tag{18}$$

This bound evaluates the common unweighted regression risk. DAR enters the analysis through the learned predictor and its attained empirical risk.

5 Experiment

5.1 Experiments Settings

Tasks and Datasets. (1) **Multimodal Crowd Counting (MCC)**: we adopted two widely used datasets: RGBT-CC Liu et al. (2021a) and DroneRGBT Peng et al. (2020). (2) **Image Fusion (IF)**. IF is commonly formulated as a regression task, since the model is required to generate a continuous-valued image. We follow TC-MoA Zhu et al. (2024) to construct the training and testing

Method	Clean				Noise			
	RGBT-CC		Drone		RGBT-CC		Drone	
	G0↓	RMSE↓	G0↓	RMSE↓	G0↓	RMSE↓	G0↓	RMSE↓
BM (ECCV'24)	10.19	17.32	6.20	10.04	34.44	58.57	24.43	32.74
PDF (ICML'24)	11.68	17.89	7.52	9.54	30.52	44.01	17.25	28.03
CRMT (ICLR'24)	10.34	20.03	5.10	8.41	34.52	49.02	10.95	17.40
EAU (CVPR'24)	18.77	32.24	12.02	17.56	30.26	52.23	17.65	28.10
Free Lunch (CVPR'25)	**9.57**	17.05	4.94	7.81	33.58	62.97	11.90	19.05
HOVQ (AAAI'26)	30.23	39.86	18.29	21.37	41.60	71.40	28.88	34.09
CPSC (CVPR'26)	15.20	30.72	22.17	30.91	60.46	94.28	14.02	21.76
SPD (Ours)	9.86	**13.04**	**4.76**	**7.24**	**22.33**	**42.82**	**9.99**	**15.83**

Dataset	$\mathcal{L}_{reg}$	DAR	Clean		Noise	
			G0↓	RMSE↓	G0↓	RMSE↓
RGBT-CC	✓	×	14.03	18.74	29.66	49.61
	×	✓	12.55	16.48	26.37	48.24
	×	×	16.35	19.76	33.50	54.28
	✓	✓	**9.86**	**13.04**	**22.33**	**42.82**
Drone	✓	×	5.88	9.41	17.05	27.22
	×	✓	5.77	8.94	17.44	27.30
	×	×	6.22	10.24	18.09	28.44
	✓	✓	**4.76**	**7.24**	**9.99**	**15.83**

Table 4: (a) Quantitative comparison on crowd counting under clean and noise settings. (b) Ablation study on crowd counting task. G0, denote Game0.

datasets. (3) **Multimodal Depth Estimation (MDE)**. We conducted experiments on MVSEC Zhu et al. (2018) and DENSE datasets Hidalgo-Carrió et al. (2020).

Settings. (1) For the MCC task, we use GAME0 and Root Mean Square Error (RMSE) as evaluation metrics. (2) For the IF task, we adopt SSIM, MS-SSIM (MS for short), QW, and FMI to assess the quality of fused images. (3) For the MDE task, we evaluate the average absolute depth errors at various cut-off distances (i.e., 10m, 20m, and 30m), including Absolute Relative Error (Abs Rel, *Abs.* for short), Logarithmic Root Mean Squared Error (RMSE log, $R_{\log}$ for short), and Scale-Invariant Logarithmic Error (SI log). In the formal paper, all noises imply Gaussian noise in the quantitative comparisons. More settings are in the appendix.

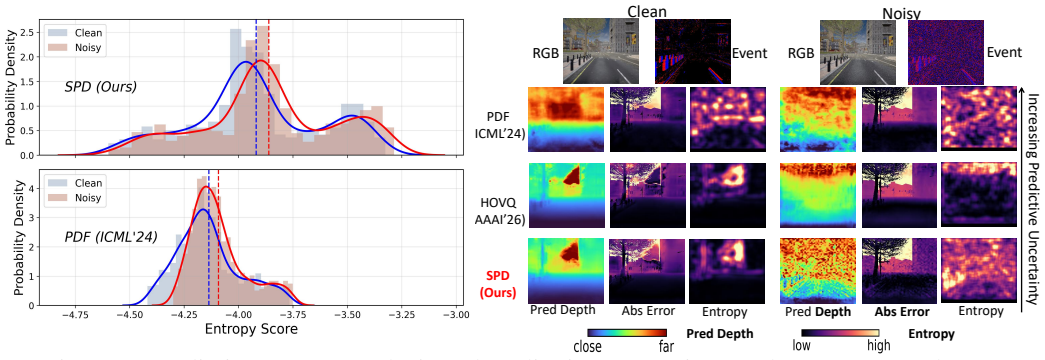

Figure 4: Predictive entropy analysis and qualitative comparison under on DENSE dataset.

5.2 Comparison with SOTA Methods

We choose the SOTA methods of each task (CCF Cao et al. (2024b) and TITA Hu et al. (2025)) for IF; BM Meng et al. (2024)) and Free-Lunch Meng et al. (2025) for MCC; ER-F2D Devulapally et al. (2024) and HAFUNet Zhang et al. (2025) for MDE task. SOTA Robust Multimodal Learning (RML) methods serve as our counterparts as well, including EAU Gao et al. (2024b), HOVQ Hu et al. (2026), PDF Cao et al. (2024a), CRMT Yang et al. (2024), and CPSC Jiang et al. (2026). **Image Fusion.** As shown in Table 1, SPD shows competitive performance under clean conditions and becomes more clearly advantageous under Gaussian noise. **Depth Estimation.** We exhibit the performance of our model on DENSE and MVSEC datasets, results are shown in Table.3 and Table.12. On DENSE, SPD exceeds CPSC by 0.023 in Abs Rel and 0.009 in RMSE log, on the other hand, with noise, it surpasses CPSC by 0.056 in Abs Rel. On MVSEC, SPD reaches sota performances as well. **Crowd Counting.** As shown in Table 4, in clean setting, our SPD achieve competitve performances compared to the task-specific method-Free Lunch on both dataset. In the noise conditions, our method surpasses both RML method and task-specific method with significant enhancement.

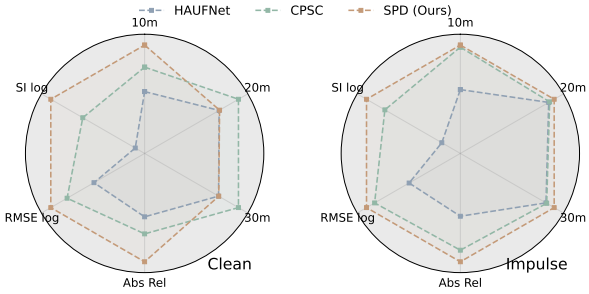

Figure 5: Performances on MVSec (Day1) with clean and impulse settings.

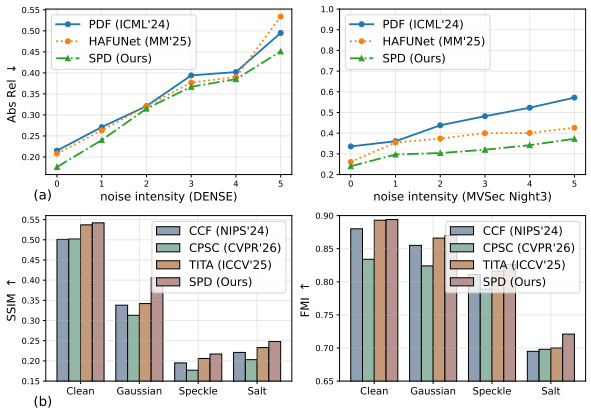

Figure 6: (a) Performance comparison with different Gaussian intensity. (b) Different noises on IF task.

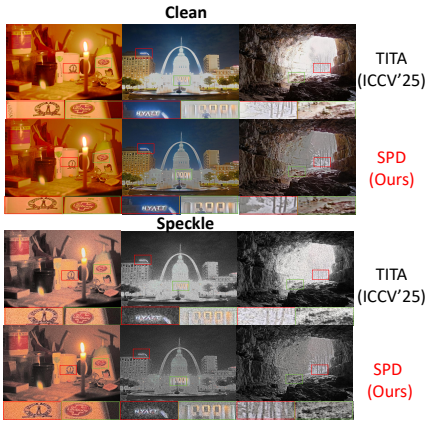

Figure 7: Visualization of image fusion under clean and noise (Speckle).

6 Further Analysis

Ablation Study. We conducted ablation study on MCC with two conditions, results are shown in right table of Table.4. The ablation results show that both modules consistently improves performance. In detail, on RGBT-CC, introducing $\mathcal{L}_{\text{reg}}$ reduces RMSE by 3.44 under clean conditions. For noisy condition, the improvement of both DAR and regularization becomes more significant.

Predictive Entropy and Robustness Visualization. Fig. 4 analyzes the predictive uncertainty and qualitative depth results under clean and noisy inputs. The left part shows that, compared with PDF, SPD exhibits a more stable entropy distribution when the input is corrupted, indicating that SPD can better preserve the reliability of its predictive distribution under degraded observations. Moreover, the qualitative results further show that compared to the counterparts, SPD preserves more coherent scene structures and produces entropy responses that better reflect degraded or ambiguous regions

Additional Results. Moreover, we also provide the performances of SPD on MVSec (Day1) under clean and Impluse noise setting. We apply an inverse normalization for visualization, where each score is computed as score $= \frac{\min(x)}{x}$. Thus, a larger radius represents better performance. In the clean setting, our method performs best on 10m, Abs Rel, RMSE log, and SI log, while remaining competitive on 20m and 300m. Under impulse noise, our method maintains a clear advantage over HAUFNet and CPSC, especially on Abs Rel, RMSE log, and SI log. Additionally, we also exhibited SPD's robustness on MDE task with different noise intensity, shown in Fig. 6. It is obvious that, SPD consistently achieves SOTA performances in the face of different noise levels. In the lower row of Fig. 6 is the results of Image Fusion under various noise settings. SPD retains substantial robustness under different noise conditions.

Visual analysis of SOTA approaches for Image Fusion. The qualitative comparison results for Image Fusion task are shown in Figure.7. Existing approaches (TITA) tend to amplify modality-specific noise, particularly in high-intensity and salient regions, leading to structural degradation and visual instability.

7 Conclusion

In this paper, we proposed **Structured Predictive Distributions (SPD)** for robust multimodal regression. Instead of relying on conventional point-wise prediction, SPD reformulates regression in a structured predictive space, where predictive representations are constrained to better control the hypothesis space. Specifically, we implement NIG parameterization to formulate the predictive space, and we also develop a distribution-aware reweighting (DAR) to improve optimization stability. Furthermore, using a common squared error risk and Rademacher complexity analysis, we establish conditional generalization-bound tightening through NIG-induced capacity control and characterize the empirical-risk condition for comparing learned predictors.

AI USE STATEMENT

We used generative AI tools for language editing and to assist in revising the theoretical framework, formulating assumptions, and drafting parts of the mathematical arguments and proofs.

REFERENCES

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in neural information processing systems*, 33:14927–14937, 2020.

Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE transactions on image processing*, 27(4):2049–2062, 2018.

Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Predictive dynamic fusion. *arXiv preprint arXiv:2406.04802*, 2024a.

Bing Cao, Xingxin Xu, Pengfei Zhu, Qilong Wang, and Qinghua Hu. Conditional controllable image fusion. *Advances in Neural Information Processing Systems*, 37:120311–120335, 2024b.

Anusha Devulapally, Md Fahim Faysal Khan, Siddharth Advani, and Vijaykrishnan Narayanan. Multi-modal fusion of event and rgb for monocular depth estimation using a unified transformer-based architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2081–2089, June 2024.

Zixian Gao, Disen Hu, Xun Jiang, Huimin Lu, Heng Tao Shen, and Xing Xu. Enhanced experts with uncertainty-aware routing for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9650–9659, 2024a.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26876–26885, 2024b.

Javier Hidalgo-Carrió, Daniel Gehrig, and Davide Scaramuzza. Learning monocular dense depth from events. In *2020 International Conference on 3D Vision (3DV)*, pp. 534–542. IEEE, 2020.

Disen Hu, Xun Jiang, Xiaofeng Cao, Zheng Wang, Jingkuan Song, Heng Tao Shen, and Xing Xu. Hyper-opinion vagueness quantification for robust multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21831–21839, 2026.

Xingyu Hu, Junjun Jiang, Chenyang Wang, Kui Jiang, Xianming Liu, and Jiayi Ma. Balancing task-invariant interaction and task-specific adaptation for unified image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11262–11272, 2025.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3496–3504, 2021.

Xun Jiang, Yufan Gu, Disen Hu, Yuqing Hou, Yazhou Yao, Fumin Shen, Heng Tao Shen, and Xing Xu. Multimodal learning on low-quality data with conformal predictive self-calibration, 2026. URL https://arxiv.org/abs/2605.03820.

Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2041–2050, 2018.

Hui Lin, Xiaopeng Hong, Zhiheng Ma, Xing Wei, Yunfeng Qiu, Yaowei Wang, and Yihong Gong. Direct measure matching for crowd counting. *arXiv preprint arXiv:2107.01558*, 2021.

Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting crowd counting via multifaceted attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19628–19637, 2022.

Hui Lin, Zhiheng Ma, Xiaopeng Hong, Qinnan Shangguan, and Deyu Meng. Gramformer: Learning crowd counting via graph-modulated transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 3395–3403, 2024.

Lingbo Liu, Jiaqi Chen, Hefeng Wu, Guanbin Li, Chenglong Li, and Liang Lin. Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4823–4833, 2021a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021b.

Huan Ma, Zongbo Han, Changqing Zhang, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. Trustworthy multimodal regression with mixture of normal-inverse gamma distributions. *Advances in Neural Information Processing Systems*, 34:6881–6893, 2021.

Nis Meinert, Jakob Gawlikowski, and Alexander Lavin. The unreasonable effectiveness of deep evidential regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9134–9142, 2023.

Haoliang Meng, Xiaopeng Hong, Chenhao Wang, Miao Shang, and Wangmeng Zuo. Multi-modal crowd counting via a broker modality. In *European conference on computer vision*, pp. 231–250. Springer, 2024.

Haoliang Meng, Xiaopeng Hong, Zhengqin Lai, and Miao Shang. Free lunch enhancements for multi-modal crowd counting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14013–14023, 2025.

Yue Ming, Xuyang Meng, Chunxiao Fan, and Hui Yu. Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438:14–33, 2021.

Mansour Nejati, Shadrokh Samavi, and Shahram Shirani. Multi-focus image fusion using dictionary-based sparse representation. *Information fusion*, 25:72–84, 2015.

Dongpin Oh and Bonggun Shin. Improving evidential deep learning via multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7895–7903, 2022.

Tao Peng, Qing Li, and Pengfei Zhu. Rgb-t crowd counting from drone: A benchmark and mmccn network. In *Proceedings of the Asian conference on computer vision*, 2020.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pp. 459–479. Springer, 2022.

Benkuan Wang, Datong Liu, Xiyuan Peng, and Zhenyu Wang. Data-driven anomaly detection of uav based on multimodal regression model. In *2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pp. 1–6. IEEE, 2019.

Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2250–2259, October 2025.

Yuefei Wu, Bin Shi, Bo Dong, Qinghua Zheng, and Hua Wei. The evidence contraction issue in deep evidential regression: Discussion and solution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21726–21734, 2024.

Shuang Xu, Xiaoli Wei, Chunxia Zhang, Junmin Liu, and Jiangshe Zhang. Mffw: A new dataset for multi-focus image fusion. *arXiv preprint arXiv:2002.04780*, 2020.

Zequn Yang, Yake Wei, Ce Liang, and Di Hu. Quantifying and enhancing multi-modal robustness with modality preference. *arXiv preprint arXiv:2402.06244*, 2024.

Kai Ye, Tiejin Chen, Hua Wei, and Liang Zhan. Uncertainty regularized evidential regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16460–16468, 2024.

Hao Zhang, Zhuliang Le, Zhenfeng Shao, Han Xu, and Jiayi Ma. Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. *Information Fusion*, 66:40–53, 2021.

Juncheng Zhang, Qingmin Liao, Shaojun Liu, Haoyu Ma, Wenming Yang, and Jing-Hao Xue. Realmff: A large realistic multi-focus image dataset with ground truth. *Pattern Recognition Letters*, 138:370–377, 2020.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pp. 41753–41769. PMLR, 2023.

Siyuan Zhang, Xiaoping Wang, Jiang Li, Weibin Feng, Xin Zhan, and Hongzhi Huang. Hafunet: A hierarchical attention fusion network for monocular depth estimation integrating event and frame data. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1395–1403, 2025.

Xingchen Zhang. Benchmarking and comparing multi-exposure image fusion algorithms. *Information Fusion*, 74:111–131, 2021.

Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018.

Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 989–997, 2019.

Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7099–7108, 2024.

A Implementation Details

A.1 Dataset Details

In the formal paper, we validate our SPD with three different tasks: Image Fusion, Multimodal Crowd Counting, and Multimodal Depth Estimation. In this section, we provide more detailed statistic of implemented datasets.

Image Fusion. We follow TC-MoA Zhu et al. (2024) to construct the training and testing datasets, using LLVIP Jia et al. (2021), SCIE Cai et al. (2018), MFI-WHU Zhang et al. (2021), and RealMFF Zhang et al. (2020) as training set and image pairs from LLVIP, MEFB Zhang (2021), Lytro Nejati et al. (2015), MFFW Xu et al. (2020), and MFI-WHU. The primary objective of image fusion is to integrate complementary information from multimodal source images into a single fused image. The fused images aims to retain the most structured-preserved and salient features from each input modality, providing a more comprehensive scene description that is robust to adverse lighting conditions. The image fusion benchmarks include infrared-visible, multi-exposure, and multi-focus image pairs. Only infrared–visible fusion involves a thermal input; the other tasks combine images captured under different exposure or focus settings.

Crowd Counting. Crowd counting aims to accurately predict the population density and total count within a scene. Multi-modal crowd counting leverages complementary information, such as RGB images and infrared images from sensors, to achieve robust performance across diverse environments. In our implementation, we utilzed the RGBT-CC Liu et al. (2021a) and DroneRGBT Peng et al. (2020), which are two widely used crowd counting datasets that are composed infrared images and RGB images. In detail, RGBT-CC serves as a benchmark designed for all-weather crowd

Task	Dataset	Type / Modality	Split / Sequence	# Samples	Usage
Image Fusion	LLVIP	Infrared-visible	Training	12,025 pairs	Training
	SCIE	Multi-exposure	Training	589 pairs	Training
	RealMFF	Multi-focus	Training	710 pairs	Training
	MFI-WHU	Multi-focus	Training	90 pairs	Training
	LLVIP	Infrared-visible	Testing	70 pairs	IVF testing
	MEFB	Multi-exposure	Testing	100 pairs	MEF testing
	Lytro	Multi-focus	Testing	20 pairs	MFF testing
	MFFW	Multi-focus	Testing	13 pairs	MFF testing
	MFI-WHU	Multi-focus	Testing	30 pairs	MFF testing
Depth Estimation	MVSEC	Real-world RGB-event	Outdoor day2	10,000 samples	Train / Val
			Outdoor night1	5,000 samples	Test
			Outdoor night2	5,000 samples	Test
			Outdoor night3	5,000 samples	Test
			Outdoor day1	5,000 samples	Test
	DENSE	Synthetic RGB-event	Town01–Town05	1,000 each	Training
			Town06–Town07	1,000 each	Validation
			Town10	1,000	Test
Crowd Counting	RGBT-CC	RGB-thermal	Train / Val / Test	1,030 / 200 / 800 pairs	–
	DroneRGBT	Drone RGB-thermal	Train / Test	3,600 pairs	–

Table 5: Dataset construction for image fusion, event-frame depth estimation, and crowd counting.

counting. It encompasses different urban scenarios captured under varying illumination conditions. DroneRGBT focuses on aerial perspectives, including scenes captured from various flight altitudes.

Depth Estimation. Depth estimation aims to perceive the distance of each pixel in a scene relative to the camera from input images. Particularly, multimodal depth estimation focuses on integrating information from multiple sensors, such as RGB cameras, Infrared sensors, or LiDAR, leveraging the complementary signals of different modalities. We utilized two widely used datasets: MVSEC Zhu et al. (2018) and DENSE datasets Hidalgo-Carrió et al. (2020). The formal provides synchronized multi-sensor recordings collected from moving platforms in both indoor and outdoor scenes. While the latter focuses on dense depth prediction with event data and complementary visual inputs.

The detailed statistical information of all the datasets is provided in the Table. 5.

A.2 Implementation Details

As illustrated in our formal paper, we utilized several different noises to each tasks according to the physical property of each input modalities. In this section, we concretely illustrate the noise implementations and our training details.

Baseline sources and comparison protocol. For image fusion, the results of TITA and CCF are reproduced using their official implementations. For depth estimation, ER-F2D and HAFUNet are reproduced using their official repositories. For crowd counting, the clean results of BM and Free Lunch are taken from prior reports, whereas the results under synthetic corruptions are obtained by us through re-evaluation. All remaining baselines are adapted by us from their official repositories to the corresponding evaluation tasks.

For controlled comparisons involving our adapted implementations, the backbone, training budget, and pretraining setup are kept consistent within each task. Results reproduced using the original implementations or taken from prior reports are distinguished from these controlled comparisons. The synthetic corruptions used in our evaluations are generated by us according to the task-specific protocols detailed below.

Image Fusion. For image fusion, we replace the NIG likelihood term with the source-image-based fusion objective adopted by TITA, comprising SSIM, texture, and intensity losses, while retaining our NIG regularization term. We optimize the model parameters using Adam with a learning rate of 2×10^{-5}. For the weighting logits introduced by FAMO, we employ a separate Adam optimizer with a learning rate of 2.5×10^{-2} and a weight decay of 1×10^{-3}. The batch size is set to 8, and the training process runs for 20,000 iterations. We evaluate Gaussian, salt-and-pepper, and speckle corruptions separately on infrared–visible, multi-exposure, and multi-focus fusion. Corruption is applied to both source images in each pair. For infrared–visible fusion, these are the infrared and

Table 6: Crowd counting corruption parameters. Gaussian standard deviations and NUC offsets are specified on the 0–255 image-intensity scale. The salt-and-pepper probability is conditional on applying the corruption. NUC σ_o denotes the initial offset-field sampling standard deviation before spatial processing.

Severity	Gaussian σ	Salt p	NUC σ_o	NUC σ_g	σ_{blur}	b
1	10	0.1	0.500	0.0000	23	64
2	20	0.2	2.375	0.0125	18	52
3	30	0.3	4.250	0.0250	13	40
4	40	0.4	6.125	0.0375	8	28
5	50	0.5	8.000	0.0500	3	16

visible images. For multi-exposure and multi-focus fusion, these are images captured under different exposure and focus settings, respectively; neither task involves a thermal modality.

For the fixed-severity comparisons, Gaussian noise is additive and zero-mean, with standard deviation $\sigma = 15$ on the 0–255 image-intensity scale. Salt-and-pepper noise has a total corruption probability of $p = 0.03$: a pixel is replaced with 0 or 255 with probability $p/2$ each. Speckle noise is generated as $I_{\text{noisy}} = I + I \odot n$, where $n \sim \mathcal{N}(0, s^2)$, $s = 0.2$, and $\odot$ denotes element-wise multiplication.

Crowd Counting. We choose Swin-T Liu et al. (2021b) as our backbone to extract features from the input modalities. We adapted Adam as our optimizer, and the momentum coefficient and temperature parameter are set to 0.999 and 0.07, respectively. The batch size is set to 16. **Crowd counting corruptions.** We evaluate Gaussian, salt-and-pepper, and non-uniformity (NUC) corruptions separately. Only the thermal input is corrupted, while the RGB input remains unchanged. Corruption is applied to the 0–255 image representation before tensor conversion and input normalization.

For each test sample, the selected corruption is applied with probability 0.5; otherwise, the thermal input is left unchanged. Thus, the reported corrupted setting uses a stochastic mixture of clean and corrupted thermal inputs, rather than corrupting every sample. This application probability is distinct from the pixel replacement probability of salt-and-pepper noise.

Let $k \in \{1, 2, 3, 4, 5\}$ denote the severity level. Gaussian noise has standard deviation $10k$. A single spatial noise field is sampled and shared across channels. For salt-and-pepper noise, the total pixel replacement probability is $p_k = k/10$, with equal probabilities $p_k/2$ of assigning 0 or 255. The fixed-severity Gaussian and salt-and-pepper comparisons use $k = 5$.

NUC is implemented as a spatially varying affine perturbation:

$$I_{\text{noisy}}(x, y) = g(x, y)I(x, y) + o(x, y). \tag{19}$$

The initial gain and offset fields are sampled as

$$g(x, y) \sim \mathcal{N}(1, \sigma_g^2), \qquad o(x, y) \sim \mathcal{N}(0, \sigma_o^2). \tag{20}$$

Writing $u = (k - 1)/4$, the implementation uses

$$\begin{aligned} \sigma_o &= 0.5 + 7.5u, \\ \sigma_g &= 0.05u, \\ \sigma_{\text{blur}} &= 20(1 - u) + 3, \\ b &= \text{int}(64 - 48u). \end{aligned} \tag{21}$$

Here, σ_{blur} and b specify the Gaussian smoothing and block-averaging parameters used to construct the offset field. Importantly, σ_o is the sampling standard deviation before these spatial operations, rather than the standard deviation of the final offset field. Table 6 summarizes the severity-dependent parameters.

Depth Estimation. We adopt pre-trained MaxViT Tu et al. (2022) as our backbone to extract the features of RGB images and Event. Our model is trained under the same experimental settings on MVSEC and DENSE. During training, we used random 224×224 cropping and random horizontal flipping for data augmentation. For event voxel grids, normalization is computed over non-zero entries, which are standardized to zero mean and unit variance; zero entries are preserved during

Table 7: Corruption severity settings for depth estimation. Gaussian standard deviations are specified on the RGB 0–255 intensity scale. Impulse ratios specify the number of location draws divided by the event voxel-grid size.

Severity	RGB Gaussian σ	Event impulse ratio r
1	10	0.05
2	20	0.15
3	30	0.25
4	40	0.50
5	50	0.80

this normalization step. We optimized our SPD with AdamW with a weight decay of 0.1. The maximum learning rate set to 3.2×10^{-5}, and the model is trained for 150 epochs with a batch size of 32. We evaluate two corruption settings separately: Gaussian noise on the RGB-frame input and impulse noise on the event voxel input.

For Gaussian corruption, independent zero-mean Gaussian perturbations are added to the RGB values on the 0–255 intensity scale. The corrupted values are clipped to $[0, 255]$ and converted to 8-bit integers before grayscale conversion and rescaling by $1/255$. The event input receives no additional corruption in this setting. Severity levels $k = 1, \ldots, 5$ correspond to standard deviations $\sigma_k = 10k$.

For impulse corruption, the non-zero event voxel values are first standardized to zero mean and unit variance when their standard deviation is positive. Corruption is then applied to the normalized voxel grid. For a grid of shape $C \times H \times W$, we draw $\lfloor r_k CHW \rfloor$ voxel locations uniformly with replacement from the entire grid, including initially zero locations. Each draw assigns a polarity of -1 or $+1$ with equal probability. The entire resulting grid is clipped to $[-1, 1]$, without subsequent standardization. The RGB input receives no additional corruption in this setting.

The impulse sampling ratios are $r_k \in \{0.05, 0.15, 0.25, 0.50, 0.80\}$ for severity levels 1–5, respectively. Since locations are sampled with replacement, r_k specifies the number of location draws relative to the grid size, rather than the exact fraction of distinct overwritten voxels.

The hyperparameter λ is set to 0.01 for all tasks. All experiments are conducted on RTX A6000 GPU.

B Detailed SPD construction

This section gives the explicit likelihood, fusion operator, regularization discrepancy, and DAR update rule used to construct and train SPD. The conditional generalization analysis for the resulting point predictors is provided in Appendix C.

Closed-form predictive likelihood. The following derivation applies to tasks using NIG likelihood supervision. Let $\hat{y}^{(i)} = \mathcal{O}(\gamma_f^{(i)})$ denote the final prediction. For each target coordinate j, we define the target-space NIG parameters as

$$\widetilde{z}_{fj}^{(i)} = \left(\hat{y}_j^{(i)}, \upsilon_{fj}^{(i)}, \alpha_{fj}^{(i)}, \beta_{fj}^{(i)}\right), \tag{22}$$

where the fused parameters are aligned with the target coordinates and satisfy $\upsilon_{fj}^{(i)} > 0$, $\alpha_{fj}^{(i)} > 1$, and $\beta_{fj}^{(i)} > 0$. The corresponding predictive density is obtained by marginalizing the latent Gaussian mean and variance under this NIG parameterization.

We use a factorized predictive likelihood over the d target coordinates:

$$p\left(y^{(i)} \mid \widetilde{z}_f^{(i)}\right) = \prod_{j=1}^{d} p\left(y_j^{(i)} \mid \widetilde{z}_{fj}^{(i)}\right). \tag{23}$$

Accordingly, the fused negative log-likelihood in Eq. 6 has the closed form

$$\begin{aligned}\mathcal{L}_{\mathrm{SPD}}^{f} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\Bigg[&\frac{1}{2}\log\left(\frac{\pi}{\upsilon_{fj}^{(i)}}\right) - \alpha_{fj}^{(i)}\log\Omega_{fj}^{(i)} \\ &+ \left(\alpha_{fj}^{(i)}+\frac{1}{2}\right)\log\Big(\upsilon_{fj}^{(i)}\left(y_j^{(i)}-\hat{y}_j^{(i)}\right)^2+\Omega_{fj}^{(i)}\Big) \\ &+ \log\frac{\Gamma\left(\alpha_{fj}^{(i)}\right)}{\Gamma\left(\alpha_{fj}^{(i)}+\frac{1}{2}\right)}\Bigg],\end{aligned} \tag{24}$$

where

$$\Omega_{fj}^{(i)} = 2\beta_{fj}^{(i)}\left(1+\upsilon_{fj}^{(i)}\right). \tag{25}$$

For scalar regression, the coordinate sum reduces to a single term.

This expression expands the fused predictive likelihood, rather than a separate modality-specific supervision loss. For this likelihood-based training setting, the objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{SPD}}^{f} + \lambda\mathcal{L}_{\mathrm{reg}}, \tag{26}$$

where $\mathcal{L}_{\mathrm{reg}}$ is defined in Eq. 7 and acts on the modality-specific representations before fusion. The target-space distribution is parameterized using the readout prediction as its location; we do not assume that applying an arbitrary readout to a random variable preserves its NIG-induced predictive family.

Complete form of Eq.4. The concrete structure-preserving operation is: Given modality-specific predictive representations $\mathbf{z}_m^{(i)} = (\gamma_m^{(i)}, \upsilon_m^{(i)}, \alpha_m^{(i)}, \beta_m^{(i)})$, the fused representation $\mathbf{z}_f^{(i)} = (\gamma_f^{(i)}, \upsilon_f^{(i)}, \alpha_f^{(i)}, \beta_f^{(i)})$ is computed by:

$$\upsilon_f^{(i)} = \sum_{m=1}^{M}\upsilon_m^{(i)}, \tag{27}$$

$$\gamma_f^{(i)} = \frac{\sum_{m=1}^{M}\upsilon_m^{(i)}\gamma_m^{(i)}}{\upsilon_f^{(i)}}, \tag{28}$$

$$\alpha_f^{(i)} = \sum_{m=1}^{M}\alpha_m^{(i)}, \tag{29}$$

$$\beta_f^{(i)} = \sum_{m=1}^{M}\beta_m^{(i)} + \frac{1}{2}\sum_{m=1}^{M}\upsilon_m^{(i)}\left(\gamma_m^{(i)}-\gamma_f^{(i)}\right)^2. \tag{30}$$

This integration operator maps modality-specific predictive representations into a unified representation while preserving the validity constraints of the structured predictive space, i.e., $\upsilon_f^{(i)} > 0$, $\alpha_f^{(i)} > 1$, and $\beta_f^{(i)} > 0$.

Complete form of Eq.7. This regularization aims to learn stable and well-constrained predictive representations by encouraging them to stay close to a fixed reference representation in the structured predictive space. Here, $(\gamma_0, \upsilon_0, \alpha_0, \beta_0)$ denotes the reference representation, where γ_0 specifies the reference center, υ_0 controls the reference concentration, and α_0, β_0 determine the reference shape and scale of the predictive representation. We instantiate it as constant tensors:

$$\gamma_0 = \mathbf{0}, \quad \upsilon_0 = 0.1\mathbf{1}, \quad \alpha_0 = 1.1\mathbf{1}, \quad \beta_0 = \mathbf{1}, \tag{31}$$

where $\mathbf{0}$ and $\mathbf{1}$ denote tensors with the same shape as the corresponding learned predictive parameters. The complete form of the regularization term is given by:

$$\begin{aligned}\mathcal{L}_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N}\Bigg[& \\ & \frac{(\alpha^{(i)}-1)v_0^{(i)}}{2\beta^{(i)}}\left(\gamma_0^{(i)}-\gamma^{(i)}\right)^2 \\ & + \frac{1}{2}\left(\frac{v_0^{(i)}}{v^{(i)}} - \log\frac{v_0^{(i)}}{v^{(i)}} - 1\right) \\ & + \alpha_0^{(i)}\log\frac{\beta^{(i)}}{\beta_0^{(i)}} - \log\frac{\Gamma\left(\alpha^{(i)}\right)}{\Gamma\left(\alpha_0^{(i)}\right)} \\ & + \left(\alpha^{(i)}-\alpha_0^{(i)}\right)\psi\left(\alpha^{(i)}\right) - \alpha^{(i)}\left(1-\frac{\beta_0^{(i)}}{\beta^{(i)}}\right)\Bigg].\end{aligned} \tag{32}$$

where $\Gamma(\cdot)$ is the Gamma function and $\psi(\cdot)$ is the digamma function. The coordinatewise discrepancy is averaged over samples, channels, and spatial locations within each representation, and then averaged over modalities and, where applicable, prediction scales. The coefficient λ multiplies this averaged regularizer.

Implementation of DAR. For depth estimation, we use the temperature-controlled weighting rule in Eq. (9). For each sample, the predictive spread is first averaged over channel and spatial dimensions. The sample coefficients and their mini-batch aggregation are

$$\begin{aligned}\bar{s}_m^{(i)} &= \text{Avg}_{C,H,W}\left(s_m^{(i)}\right), \\ w_m^{(i)} &= \exp\left(-\frac{\bar{s}_m^{(i)}}{T}\right), \qquad T>0, \\ \bar{w}_m &= \frac{1}{n}\sum_{i=1}^{n}\text{stopgrad}\left(w_m^{(i)}\right).\end{aligned} \tag{33}$$

Here, n is the mini-batch size. For numerical stability, the denominators of the two spread terms are evaluated as $\alpha - 1 + \epsilon$ and $v(\alpha-1)+\epsilon$, respectively, with $\epsilon = 10^{-6}$.

After backpropagating the total training objective $\mathcal{L}_B$, including task supervision and NIG regularization, we rescale the head gradients:

$$g_m^{\text{DAR}} = \bar{w}_m \nabla_{\theta_m}\mathcal{L}_B. \tag{34}$$

Here, θ_m contains the convolutional weights of the modality-specific NIG prediction head. Head biases, shared fusion modules, and the final readout retain their ordinary backpropagated gradients before global gradient clipping. The same modulation is applied independently at each prediction scale.

The modulation coefficients are recomputed from the current predictive representations at every training step and detached before gradient modulation. For sequence inputs, the coefficients are computed from the final forward pass, while the total training loss aggregates the sequence losses. After head-gradient modulation, the global gradient norm is clipped to 0.1 before the optimizer step. **Rationale and scope of DAR.** For a scalar NIG representation z_m, let $\widetilde{Y}$ follow its induced predictive distribution. The spread satisfies

$$s_m = \text{Var}(\widetilde{Y} \mid z_m) = \mathbb{E}\Big[(\widetilde{Y}-\gamma_m)^2 \mid z_m\Big]. \tag{35}$$

Thus, predictive spread quantifies the model implied expected squared deviation around the predictive mean and motivates its use as an uncertainty signal for gradient modulation. Its association with observed prediction errors depends on the quality of the learned predictive representation; large spread can arise from both corrupted inputs and difficult clean examples. In our implementation, the modulation coefficients are recomputed from the current predictive representations at each update, but are detached from the computational graph and treated as constants within that update. Consequently, no derivatives through these coefficients are propagated.

C Complete Theoretical Analysis

C.1 Common risk and scope of the comparison

The theory evaluates the final point prediction using the common loss $\ell(h, x, y) = \|h(x) - y\|_2^2$ from Definition 1. The NIG likelihood and DAR specify training, whereas $\mathcal{R}_\ell(h)$ and $\widehat{\mathcal{R}}_\ell(h)$ evaluate the resulting point predictor. The comparison does not require a learned predictor to minimize the squared error training objective exactly.

A point predictor h may also be regarded as the mean of $\mathcal{N}(h(x), \sigma_0^2 I_d)$ with fixed $\sigma_0^2 > 0$. Its negative log-likelihood equals:

$$-\log p_G(y \mid h(x)) = \frac{\|h(x) - y\|_2^2}{2\sigma_0^2} + \frac{d}{2}\log(2\pi\sigma_0^2). \tag{36}$$

This identity explains the Gaussian interpretation of the reference point-prediction class. All risk comparisons below nevertheless use the same squared error loss.

For a vector-valued class $\mathcal{H}$, write $\mathcal{H}_j = \{x \mapsto h_j(x) : h \in \mathcal{H}\}$ and $\mathfrak{C}_{\mathcal{S}}(\mathcal{H}) = \sum_{j=1}^d \widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{H}_j)$, as in Lemma 1.

C.2 Assumptions and matched prediction classes

Assumption 1 (NIG regularization, matched architecture, and bounded regression domain). Let $\mathcal{S} \sim \mathcal{D}^N$ be the sample used for the risk bound. Let Q be an input reference measure fixed independently of $\mathcal{S}$. It may be the empirical measure of a separate fixed reference set. All parameter domains and constants below are specified independently of $\mathcal{S}$; the analysis can be conditioned on an independently chosen reference set.

Scope of the analysis. Assumption 1 specifies a constrained model family for which NIG regularization yields explicit capacity control. Accordingly, the results characterize conditional generalization bound tightening within the analyzed family.

Location heads and feature conditions. There are M modalities, each with a q-dimensional NIG location:

$$\gamma_m(x) = W_m \phi_{m,\theta}(x), \qquad W_m \in \mathbb{R}^{q \times p_m}, \quad \theta \in \Theta. \tag{37}$$

Here x contains all input modalities, allowing earlier cross-modal feature interactions. The feature maps do not use the location-head matrices $W = (W_1, \ldots, W_M)$. For constants $B < \infty$ and $\kappa > 0$, every admissible θ and modality satisfy:

$$\begin{aligned} \sup_{x \in \mathcal{X}} \|\phi_{m,\theta}(x)\|_2 &\le B, \\ \Sigma_{m,\theta} := \mathbb{E}_Q[\phi_{m,\theta}(X)\phi_{m,\theta}(X)^\top] &\succeq \kappa I. \end{aligned} \tag{38}$$

These conditions are uniform over Θ. An affine location head can be represented by augmenting its feature vector with a constant coordinate; the same Gram condition is then required for that augmented feature vector.

Fusion and readout. For coordinate j, the fusion weights are:

$$a_{mj,\theta}(x) = \frac{\upsilon_{mj,\theta}(x)}{\sum_{k=1}^M \upsilon_{kj,\theta}(x)}, \qquad \upsilon_{mj,\theta}(x) > 0. \tag{39}$$

Thus $a_{mj,\theta} \ge 0$ and $\sum_m a_{mj,\theta} = 1$. Writing $A_{m,\theta}(x) = \mathrm{diag}(a_{m1,\theta}, \ldots, a_{mq,\theta})$, the final point prediction is:

$$h_{W,\theta}(x) = O_\theta \sum_{m=1}^M A_{m,\theta}(x) W_m \phi_{m,\theta}(x), \tag{40}$$

where $\|O_\theta\|_{\mathrm{op}} \le K$ and $O_\theta \in \mathbb{R}^{d \times q}$ is linear, has no additional free additive bias, and $K < \infty$. For fixed θ, the feature maps, fusion weights, and readout do not depend on W. This is architectural parameter separation, not statistical independence of fitted parameters. The domain Θ is the same for all comparison radii.

Dataset	Method	Clean				Gaussian			
		SSIM↑	MS↑	QW↑	FMI↑	SSIM↑	MS↑	QW↑	FMI↑
LLVIP	CCF (NIPS'24)	0.458	0.921	0.726	0.881	0.364	0.784	0.814	0.861
	EAU (CVPR'24)	0.450	0.815	0.846	0.822	0.378	0.781	0.786	0.839
	PDF (ICML'24)	0.458	0.905	0.861	0.847	0.384	0.792	0.811	0.820
	CRMT (ICLR'24)	0.449	0.909	0.878	0.871	0.381	0.814	0.813	0.869
	TITA (ICCV'25)	0.468	0.936	0.889	**0.896**	0.372	0.809	0.828	0.878
	HOVQ (AAAI'26)	0.356	0.741	0.723	0.800	0.299	0.675	0.763	0.791
	CPSC (CVPR'26)	0.414	0.911	0.848	0.835	0.309	0.755	0.785	0.815
	SPD (Ours)	**0.469**	**0.938**	**0.890**	**0.896**	**0.395**	**0.820**	**0.830**	**0.879**
MEF	CCF (NIPS'24)	0.421	0.949	0.820	0.885	0.324	0.811	0.721	0.851
	EAU (CVPR'24)	0.425	0.952	0.815	0.894	0.314	0.795	0.716	0.837
	PDF (ICML'24)	0.437	0.928	0.802	0.881	0.320	0.787	0.720	0.844
	CRMT (ICLR'24)	0.430	0.954	0.802	0.886	0.329	0.818	0.718	0.852
	TITA (ICCV'25)	0.437	0.960	0.813	0.900	0.336	0.822	0.728	0.862
	HOVQ (AAAI'26)	0.378	0.857	0.761	0.811	0.294	0.764	0.685	0.810
	CPSC (CVPR'26)	0.422	0.944	0.796	0.823	0.313	0.797	0.721	0.836
	SPD (Ours)	**0.443**	**0.961**	**0.829**	**0.901**	**0.342**	**0.830**	**0.741**	**0.866**
MFF	CCF (NIPS'24)	0.625	0.952	0.651	0.875	0.325	0.846	0.824	0.855
	EAU (CVPR'24)	0.674	0.966	0.843	0.870	0.322	0.843	0.820	0.847
	PDF (ICML'24)	0.632	0.960	0.912	0.865	0.317	0.838	0.816	0.848
	CRMT (ICLR'24)	0.691	0.974	0.926	0.879	0.331	0.825	0.803	0.852
	TITA (ICCV'25)	0.706	0.992	**0.932**	**0.885**	0.319	0.854	0.832	0.860
	HOVQ (AAAI'26)	0.622	0.887	0.669	0.832	0.286	0.785	0.800	0.807
	CPSC (CVPR'26)	0.669	0.987	0.884	0.846	0.318	0.802	0.815	0.822
	SPD (Ours)	**0.714**	**0.994**	0.927	**0.885**	**0.482**	**0.858**	**0.834**	**0.861**

Table 8: Quantitative comparison on image fusion datasets under clean and Gaussian noise settings.

Bounded NIG aleatoric variance. Use common fixed reference parameters $z_0 = (0, \upsilon_0, \alpha_0, \beta_0)$ with $\upsilon_0 > 0$, $\alpha_0 > 1$, and $\beta_0 > 0$. For every NIG coordinate:

$$0 < V_{mj}(x) := \frac{\beta_{mj}(x)}{\alpha_{mj}(x) - 1} \leq \overline{V} < \infty \tag{41}$$

for Q-almost every x.

Explicit NIG regularization budget. Let D be the coordinatewise discrepancy in the stated NIG regularizer. The constrained SPD class consists of models satisfying:

$$\Omega_Q(z) := \frac{1}{Mq} \sum_{m=1}^{M} \sum_{j=1}^{q} \mathbb{E}_Q[D(z_{mj}(X), z_0)] \leq \tau, \tag{42}$$

for a fixed $\tau > 0$. Assume this feasible class is nonempty. This is a constraint defining the analyzed class. The NIG coordinate dimension q need not equal the final output dimension d. The displayed normalization averages over both modalities and NIG coordinates. If coordinates are summed instead, remove q from the normalization and from the radius derived below.

Reference class and bounded targets. Let:

$$\mathcal{H}(t) = \{h_{W,\theta} : \theta \in \Theta,\ \|W\|_F \leq t\}, \tag{43}$$

where $\|W\|_F^2 = \sum_{m=1}^{M} \|W_m\|_F^2$. For a fixed $R > 0$, define $\mathcal{H}_G^{\text{point}}(R) = \mathcal{H}(R)$. Its members can be the means of fixed-variance Gaussian predictors; Let $\mathcal{H}_{\text{SPD}}^{\text{point}}$ be all final point predictions admitting NIG parameters that satisfy the preceding conditions, without imposing a separate norm bound on W. Assume $\|Y\|_2 \leq B_y < \infty$ almost surely under $\mathcal{D}$.

C.3 NIG-INDUCED CLASS INCLUSION

Proposition C.1 (NIG-induced class inclusion). Under Assumption 1, define:

$$r_{\mathrm{NIG}} = \sqrt{\frac{2Mq\overline{V}\tau}{\upsilon_0\kappa}}. \tag{44}$$

Then:

$$\mathcal{H}_{\mathrm{SPD}}^{\mathrm{point}} \subseteq \mathcal{H}(r_{\mathrm{NIG}}). \tag{45}$$

In particular, if:

$$\tau \le \frac{\upsilon_0\kappa R^2}{2Mq\overline{V}}, \tag{46}$$

then $r_{\mathrm{NIG}} \le R$ and:

$$\mathcal{H}_{\mathrm{SPD}}^{\mathrm{point}} \subseteq \mathcal{H}(r_{\mathrm{NIG}}) \subseteq \mathcal{H}_G^{\mathrm{point}}(R). \tag{47}$$

Proof. With reference location zero, the discrepancy in the paper decomposes as:

$$D(z, z_0) = \frac{(\alpha-1)\upsilon_0}{2\beta}\gamma^2 + A(\upsilon, \alpha, \beta), \tag{48}$$

$$\begin{aligned} A(\upsilon, \alpha, \beta) &= \frac{1}{2}\left(\frac{\upsilon_0}{\upsilon} - \log\frac{\upsilon_0}{\upsilon} - 1\right) \\ &\quad + D_{\mathrm{KL}}\left(\mathrm{IG}(\alpha, \beta) \,\|\, \mathrm{IG}(\alpha_0, \beta_0)\right) \ge 0. \end{aligned} \tag{49}$$

The inverse-gamma distribution uses the shape–scale convention. The two terms in A are nonnegative because $u - \log u - 1 \ge 0$ for $u > 0$ and KL divergence is nonnegative. This decomposition concerns the stated discrepancy. The variance cap therefore implies:

$$D(z_{mj}(x), z_0) \ge \frac{\upsilon_0}{2V_{mj}(x)}\gamma_{mj}(x)^2 \ge \frac{\upsilon_0}{2\overline{V}}\gamma_{mj}(x)^2 \tag{50}$$

Q-almost everywhere. Summing the coordinates and modalities gives:

$$\begin{aligned} \tau \ge \Omega_Q(z) &\ge \frac{\upsilon_0}{2Mq\overline{V}} \sum_m \mathbb{E}_Q[\|W_m\phi_{m,\theta}(X)\|_2^2] \\ &= \frac{\upsilon_0}{2Mq\overline{V}} \sum_m \mathrm{tr}(W_m\Sigma_{m,\theta}W_m^\top) \\ &\ge \frac{\upsilon_0\kappa}{2Mq\overline{V}} \sum_m \|W_m\|_F^2. \end{aligned} \tag{51}$$

Thus $\|W\|_F \le r_{\mathrm{NIG}}$, proving the first inclusion. Condition (46) gives the second.

C.4 CAPACITY COMPARISON FOR VECTOR-VALUED PREDICTIONS

Proposition C.2 (Capacity contraction). Let $\rho = r_{\mathrm{NIG}}/R \le 1$ and use the coordinate-complexity functional $\mathfrak{C}_{\mathcal{S}}$ defined in the main text. For every sample,

$$\mathfrak{C}_{\mathcal{S}}(\mathcal{H}_{\mathrm{SPD}}^{\mathrm{point}}) \le \rho\mathfrak{C}_{\mathcal{S}}(\mathcal{H}_G^{\mathrm{point}}(R)). \tag{52}$$

Proof. The assumed parameter separation and linear readout imply $h_{sW,\theta} = sh_{W,\theta}$ for every $s > 0$. Consequently, as sets of vector-valued functions,

$$\mathcal{H}(r_{\mathrm{NIG}}) = \rho\mathcal{H}(R). \tag{53}$$

The same scaling holds for every scalar coordinate class. Applying monotonicity and positive homogeneity of scalar empirical Rademacher complexity coordinatewise, and summing, yields

$$\begin{aligned} \mathfrak{C}_{\mathcal{S}}(\mathcal{H}_{\mathrm{SPD}}^{\mathrm{point}}) &\le \mathfrak{C}_{\mathcal{S}}(\mathcal{H}(r_{\mathrm{NIG}})) \\ &= \rho\mathfrak{C}_{\mathcal{S}}(\mathcal{H}(R)). \end{aligned} \tag{54}$$

These quantities are finite. Indeed, coordinatewise convexity gives

$$\begin{aligned} \left\|\sum_m A_{m,\theta}\gamma_m\right\|_2^2 &\le \sum_{j=1}^q \sum_m a_{mj,\theta}\gamma_{mj}^2 \\ &\le \sum_m \|\gamma_m\|_2^2 \le B^2\|W\|_F^2. \end{aligned} \tag{55}$$

Hence $\|h_{W,\theta}(x)\|_2 \le KB\|W\|_F$.

Dataset	Method	Speckle				Salt			
		SSIM↑	MS↑	QW↑	FMI↑	SSIM↑	MS↑	QW↑	FMI↑
LLVIP	CCF (NIPS'24)	0.221	0.781	0.788	0.833	0.178	0.664	0.721	0.697
	EAU (CVPR'24)	0.226	0.778	0.774	0.825	0.176	0.659	0.718	0.682
	PDF (ICML'24)	0.230	0.772	0.768	0.822	0.180	0.672	0.736	0.687
	CRMT (ICLR'24)	0.233	0.782	0.801	0.842	0.197	0.693	0.752	0.712
	TITA (ICCV'25)	0.234	0.780	0.796	0.836	0.189	0.674	0.742	0.697
	HOVQ (AAAI'26)	0.188	0.745	0.781	0.782	0.153	0.630	0.651	0.583
	CPSC (CVPR'26)	0.205	0.765	0.778	0.804	0.164	0.684	0.646	0.715
	SPD (Ours)	**0.240**	**0.792**	**0.804**	**0.850**	**0.217**	**0.726**	**0.778**	**0.740**
MEF	CCF (NIPS'24)	0.114	0.615	0.554	0.795	0.182	0.683	0.656	0.690
	EAU (CVPR'24)	0.101	0.602	0.550	0.787	0.179	0.677	0.633	0.674
	PDF (ICML'24)	0.119	0.620	0.582	0.798	0.188	0.685	0.667	0.690
	CRMT (ICLR'24)	**0.125**	0.622	0.580	0.805	0.192	0.695	0.682	0.689
	TITA (ICCV'25)	0.119	0.618	0.577	0.806	0.197	0.698	0.680	0.693
	HOVQ (AAAI'26)	0.098	0.577	0.511	0.762	0.170	0.648	0.616	0.669
	CPSC (CVPR'26)	0.115	0.607	0.510	0.774	0.177	0.655	0.624	0.682
	SPD (Ours)	**0.125**	**0.636**	**0.592**	**0.809**	**0.198**	**0.699**	**0.696**	**0.701**
MFF	CCF (NIPS'24)	0.252	0.721	0.711	0.804	0.305	0.761	0.736	0.700
	EAU (CVPR'24)	0.245	0.711	0.695	0.793	0.300	0.752	0.750	0.709
	PDF (ICML'24)	0.258	0.730	0.714	0.810	0.312	0.749	0.755	0.701
	CRMT (ICLR'24)	0.271	0.752	0.720	0.812	0.320	0.762	0.753	0.715
	TITA (ICCV'25)	0.267	0.747	0.718	0.807	0.313	0.755	0.749	0.711
	HOVQ (AAAI'26)	0.194	0.689	0.703	0.774	0.241	0.709	0.714	0.684
	CPSC (CVPR'26)	0.210	0.712	0.723	0.789	0.269	0.717	0.735	0.697
	SPD (Ours)	**0.288**	**0.771**	**0.745**	**0.819**	**0.331**	**0.770**	**0.777**	**0.724**

Table 9: Quantitative comparison on image fusion datasets under Speckle and Salt noise settings.

C.5 Proof of Lemma 1 and Corollary 1

Lemma C.1 (Common loss and vector-output contraction). Assume $r_{\mathrm{NIG}} \le R$ and set

$$B_h = KBR, \qquad B_\ell = (B_y + B_h)^2, \qquad L = 2(B_y + B_h). \tag{56}$$

For either comparison class, the common squared error loss satisfies $0 \le \ell(h, x, y) \le B_\ell$ and

$$\widehat{\mathfrak{R}}_{\mathcal{S}}(\ell \circ \mathcal{H}) \le L\mathfrak{C}_{\mathcal{S}}(\mathcal{H}). \tag{57}$$

Proof. The output bound follows from Proposition C.2, and $\|h(x) - y\|_2 \le B_h + B_y$ gives the common loss range. For every fixed training label and coordinate, the scalar function $t \mapsto (t - y_j^{(i)})^2$ is L-Lipschitz on $[-B_h, B_h]$. Using subadditivity of the supremum and scalar contraction,

$$\begin{aligned}
\widehat{\mathfrak{R}}_{\mathcal{S}}(\ell \circ \mathcal{H}) &= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{h \in \mathcal{H}} \frac{1}{N} \sum_i \sigma_i \sum_{j=1}^{d} (h_j(x^{(i)}) - y_j^{(i)})^2 \\
&\le \sum_{j=1}^{d} \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f \in \mathcal{H}_j} \frac{1}{N} \sum_i \sigma_i (f(x^{(i)}) - y_j^{(i)})^2 \\
&\le L \sum_{j=1}^{d} \widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{H}_j).
\end{aligned} \tag{58}$$

Centering each scalar loss by its value at zero, if needed for the contraction statement, leaves its empirical Rademacher complexity unchanged. This establishes the vector-output result without applying a scalar predictor formula to a vector-valued function.

Proof of Lemma 1 and Corollary 1. For a fixed scalar loss class with range $[0, B_\ell]$, the standard empirical Rademacher bound gives, with probability at least $1 - \eta$,

$$\mathcal{R}_\ell(h) \le \widehat{\mathcal{R}}_\ell(h) + 2\widehat{\mathfrak{R}}_{\mathcal{S}}(\ell \circ \mathcal{H}) + 3B_\ell \sqrt{\frac{\log(2/\eta)}{2N}} \tag{59}$$

uniformly over $h \in \mathcal{H}$. For completeness, define $Z(\mathcal{S}) = \sup_f(\mathbb{E}[f] - N^{-1}\sum_i f(Z_i))$. Symmetrization gives $\mathbb{E}[Z] \le 2\mathbb{E}[\widehat{\mathfrak{R}}_{\mathcal{S}}]$. Both Z and the empirical Rademacher complexity change by at most B_ℓ/N when one example is replaced. Two bounded-difference inequalities, each with failure probability $\eta/2$, yield the displayed confidence term.

Apply this result with $\eta = \delta/2$ separately to $\mathcal{H}(r_{\mathrm{NIG}})$ and $\mathcal{H}(R)$ and take a union bound. Use the same constants in both bounds:

$$C_{N,\delta} = 3B_\ell\sqrt{\frac{\log(4/\delta)}{2N}}. \tag{60}$$

Lemma C.1 proves the main-text Lemma 1 on this common event. Define $\rho = r_{\mathrm{NIG}}/R$ and $\Gamma_G = 2L\mathfrak{C}_{\mathcal{S}}(\mathcal{H}_G^{\mathrm{point}}(R))$, as in Corollary 1. For every feasible SPD predictor, Proposition C.1 and $\mathfrak{C}_{\mathcal{S}}(\mathcal{H}(r_{\mathrm{NIG}})) = \rho\mathfrak{C}_{\mathcal{S}}(\mathcal{H}(R))$ then imply:

$$\mathcal{R}_\ell(h) \le \widehat{\mathcal{R}}_\ell(h) + \rho\Gamma_G + C_{N,\delta} = U_{\mathrm{NIG}}(h). \tag{61}$$

The same h also belongs to the reference class and has the valid bound $U_G(h) = \widehat{\mathcal{R}}_\ell(h)+\Gamma_G+C_{N,\delta}$. Therefore:

$$U_G(h) - U_{\mathrm{NIG}}(h) = (1-\rho)\Gamma_G \ge 0. \tag{62}$$

This proves tightening relative to the reference-class bound for the same predictor, with strict improvement when $\rho < 1$ and $\Gamma_G > 0$.

C.6 Empirical risk and comparison across predictors

Proposition C.3 (Empirical fit and cross-predictor bounds). Let $h_{\mathrm{SPD}} \in \mathcal{H}_{\mathrm{SPD}}^{\mathrm{point}}$ and $h_G \in \mathcal{H}_G^{\mathrm{point}}(R)$ be any predictors, including data-dependent fitted predictors. Define:

$$\Delta_{\mathrm{emp}} = \widehat{\mathcal{R}}_\ell(h_{\mathrm{SPD}}) - \widehat{\mathcal{R}}_\ell(h_G). \tag{63}$$

Then:

$$U_{\mathrm{NIG}}(h_{\mathrm{SPD}}) - U_G(h_G) = \Delta_{\mathrm{emp}} - (1-\rho)\Gamma_G. \tag{64}$$

Thus $\Delta_{\mathrm{emp}} \le (1-\rho)\Gamma_G$ is exactly the condition for ordering these two stated upper bounds.

Proof and interpretation. Subtract the two upper-bound expressions. Their common confidence terms cancel, giving Eq. (64). The empirical term can also be written explicitly. Set $e_i = h_G(x^{(i)}) - y^{(i)}$ and $d_i = h_{\mathrm{SPD}}(x^{(i)}) - h_G(x^{(i)})$. Then

$$\Delta_{\mathrm{emp}} = \frac{1}{N}\sum_i \left(2\langle e_i, d_i\rangle + \|d_i\|_2^2\right). \tag{65}$$

In particular, if $D_{\mathcal{S}} = N^{-1}\sum_i \|d_i\|_2^2$, Cauchy–Schwarz gives

$$\Delta_{\mathrm{emp}} \le 2\sqrt{\widehat{\mathcal{R}}_\ell(h_G)D_{\mathcal{S}}} + D_{\mathcal{S}}. \tag{66}$$

Neither this identity nor the bound comparison requires global optimization of the squared error training objective. The actual gap, rather than a distribution-family limit, is the quantity relevant to comparing fitted predictors.

C.7 Fusion and distribution-aware reweighting

The analyzed output is the final point prediction $h(x) = O_\theta\sum_m A_{m,\theta}(x)W_m\phi_{m,\theta}(x)$. The precision-weighted mean fusion connects this expression to the construction in the method section, subject to the readout conditions in Assumption 1. Its empirical evaluation risk remains $N^{-1}\sum_i \|h(x^{(i)}) - y^{(i)}\|_2^2$.

DAR specifies a modulated update direction of the form

$$g_m^{\mathrm{DAR}} = \bar{w}_m\nabla_{\theta_m}\mathcal{L}_{\mathcal{B}}. \tag{67}$$

DAR uses the detached gradient modulation specified in Appendix C. The generalization analysis concerns the resulting predictor under the stated class conditions.

Consequently, it applies to the output h_{DAR} of a data-dependent DAR training procedure whenever $h_{\mathrm{DAR}} \in \mathcal{H}_{\mathrm{SPD}}^{\mathrm{point}}$:

$$\mathcal{R}_\ell(h_{\mathrm{DAR}}) \le \widehat{\mathcal{R}}_\ell(h_{\mathrm{DAR}}) + \rho\Gamma_G + C_{N,\delta}. \tag{68}$$

Comparison with a reference predictor uses the empirical-risk gap in Section C.6. The condition $0 < w_m^{(i)} \le 1$ alone establishes neither an additional capacity contraction for model-dependent weights nor an ordering of unweighted population risks.

C.8 Empirical bound proxy and its theoretical connection

Common diagnostic. At noise level s, let $\mathcal{S}_s$ contain N_s observations. Write $e_{m,s,i,j} = h_{m,j}(x_s^{(i)}) - y_j^{(i)}$. Let $v_{s,i,j} \in \{0,1\}$ be a validity mask determined independently of the compared predictor, and use the same evaluated observations, mask, units, and target scaling for all methods. There are $n_s = \sum_{i=1}^{N_s}\sum_{j=1}^{d} v_{s,i,j} > 0$ valid scalar targets. The common diagnostic is

$$\begin{aligned} E_{m,s} &= \frac{1}{n_s}\sum_{i,j} v_{s,i,j} e_{m,s,i,j}^2, \\ A_{m,s} &= \frac{1}{n_s}\sum_{i,j} v_{s,i,j} |e_{m,s,i,j}|, \\ P_{m,s} &= E_{m,s} + 2C_0 A_{m,s}, \qquad C_0 = 1. \end{aligned} \tag{69}$$

Thus $P_{m,s}$ combines pooled MSE and MAE; it uses unweighted prediction errors for every method, including predictors trained with DAR. The fixed coefficient C_0 is a diagnostic weight, not a measured Rademacher complexity. The absolute-error statistic is not a uniform Lipschitz constant. The interpretation of Fig. 3 requires its plotted values to be computed by this common protocol; mixing MSE, NLL, or method-dependent weights across the compared methods is not covered.

Signed changes. The quantity labeled "Bound Sensitivity" is interpreted as

$$\Delta P_{m,s} = P_{m,s} - P_{m,s-1}, \qquad s \ge 2. \tag{70}$$

It is a signed change between adjacent discrete severity levels. Negative values mean that the score decreased; an absolute sensitivity comparison would instead concern $|\Delta P_{m,s}|$.

Normalization and a conservative risk bound. The main-text squared loss sums output coordinates, whereas $E_{m,s}$ averages valid scalar targets. Their aggregation factors must be retained. If all targets are valid, $n_s = N_s d$ and $\widehat{\mathcal{R}}_\ell(h_m) = dE_{m,s}$. For label- or input-defined validity masks, define the evaluated loss

$$\ell_v(h,x,y) = \sum_{j=1}^{d} v_j(x,y)(h_j(x) - y_j)^2, \qquad \mathcal{R}_{v,s}(h) = \mathbb{E}_{\mathcal{D}_s}[\ell_v(h,X,Y)]. \tag{71}$$

The mask does not depend on h, and the same bounded-loss and coordinate-contraction proof applies: each scalar loss has derivative bounded by L, and $0 \le \ell_v \le B_\ell$. For all-valid targets this is exactly the risk in Definition 1. Set $\zeta_s = n_s/N_s$. Then

$$\widehat{\mathcal{R}}_{v,s}(h_m) = \zeta_s E_{m,s} \le \zeta_s P_{m,s}. \tag{72}$$

Here N_s is the number of observations satisfying the sampling assumption, not a count of correlated pixels within the same image. Pooled pixel averages and unweighted averages of per-image MSEs are different aggregation rules when valid-pixel counts differ; the displayed identity specifically uses the pooled convention above.

For a fixed severity, apply the analysis to $\mathcal{S}_s \sim \mathcal{D}_s^{N_s}$ with all required classes and bounds fixed independently of that sample. Write $\Gamma_{G,s} = 2L\mathfrak{C}_{\mathcal{S}_s}(\mathcal{H}_G^{\mathrm{point}}(R))$ and $C_s = 3B_\ell\sqrt{\log(4/\delta)/(2N_s)}$. On the resulting common event of probability at least $1-\delta$, for predictors in their respective classes,

$$\begin{aligned} \mathcal{R}_{v,s}(h_{\mathrm{SPD}}) &\le \zeta_s P_{\mathrm{SPD},s} + \rho\Gamma_{G,s} + C_s =: \widetilde{U}_{\mathrm{SPD},s}, \\ \mathcal{R}_{v,s}(h_G) &\le \zeta_s P_{G,s} + \Gamma_{G,s} + C_s =: \widetilde{U}_{G,s}. \end{aligned} \tag{73}$$

Their difference is:

$$\widetilde{U}_{\mathrm{SPD},s} - \widetilde{U}_{G,s} = \zeta_s(P_{\mathrm{SPD},s} - P_{G,s}) - (1-\rho)\Gamma_{G,s}. \tag{74}$$

Consequently, under the stated class conditions, $P_{\mathrm{SPD},s} \le P_{G,s}$ and $\rho \le 1$ imply an ordering of these two conservative upper bounds. Replacing empirical risk by $\zeta_s P_{m,s}$ adds a nonnegative term, so these bounds are weaker than their counterparts using the actual empirical squared error. The criterion based on Δ_{emp} in Section C.6 continues to govern the original bounds.

Applicability and relation to the model. This proposition characterizes capacity control within a constrained class induced by the NIG formulation. Location heads correspond to mean-output branches; D and the zero reference location are as in the NIG regularizer. Affine heads absorb bias via augmented features; vector locations require matrix formulation and consistent regularizer normalization; fusion follows the precision-weighted mean rule. The analysis additionally assumes bounded aleatoric variance, a reference-measure regularization budget, uniform feature bounds and nondegeneracy, and a bounded homogeneous readout. The reference measure Q is independent of the generalization sample. The Gaussian reference class shares the feature, fusion, and readout families but need not represent all experimental baselines. Thus the result is capacity control within this constrained class; implementation requires verifying the conditions and matching the fusion and readout operators.

D Additional Experimental Results

Due to the space limitation, we failed to include the complete experiment results in our formal paper with different kinds of noises. In this section, we will exhibit the complete results to show the robustness of our SPD.

D.1 Complete Experimental Results

The complete results of our SPD is shown in the Table. 8,9,12,11. It is clear to see that our SPD achieves sota performances across all the tasks and datasets and exhibited competitve robustness against various kinds of noises.

Metric	HAF.	CPSC	HOVQ	Ours
GFLOPs	14.97	14.56	14.65	14.97
Params (M)	79.9	84.1	77.1	79.9
Time (s)	0.0397	0.0404	0.0421	0.0397
FPS	25.17	24.73	23.73	25.17
Mem. (MB)	371.48	466.50	340.24	371.48

Table 10: cost comparison on DENSE.

D.2 Further Analysis

Visualization results of Depth Estimation. We output the results of predicted depth image of HAFUNet and SPD, results are shown in Fig 8. Under clean conditions, our SPD model produces more coherent depth maps and achieves lower absolute errors than HAFUNet. More importantly, when the event modality is corrupted by severe impulse noise, the baseline predictions become noticeably degraded. In contrast, SPD better preserves the global road layout, object boundaries, and distant structures, reducing the absolute error from 2.31 to 1.78 and from 2.56 to 1.95 in the two noisy examples. These results further demonstrate the robustness of the proposed structured predictive space.

Computational Effectiveness. Table 10 reports the computational cost of different methods on the DENSE benchmark. Compared with CPSC and HOVQ, our method achieves competitive or better efficiency. Although CPSC and HOVQ have slightly lower GLOPs, they do not translate into faster inference. Our method obtains the fastest inference speed together with HAFUNet, while also achieving the highest FPS among all compared methods. In addition, our GPU memory consumption is substantially lower than CPSC and only moderately higher than HOVQ. These results indicate that the proposed SPD framework improves robustness and predictive reliability without introducing extra computational overhead.

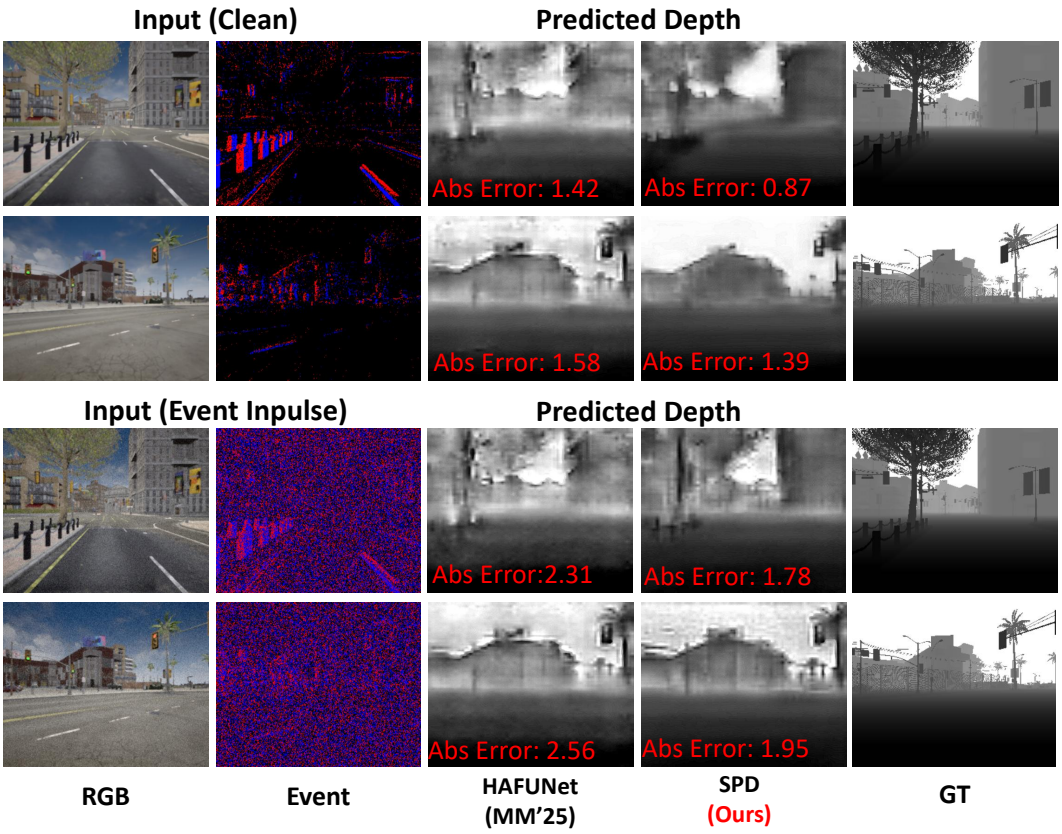

Figure 8: Depth estimation results on the DENSE dataset.

Table 11: Quantitative comparison on crowd counting under clean and different noisy settings. G0 denotes Game0.

Method	Clean				Gaussian				NUC				Salt			
	RGBT-CC		Drone		RGBT-CC		Drone		RGBT-CC		Drone		RGBT-CC		Drone	
	G0↓	RMSE↓	G0↓	RMSE↓	G0↓	RMSE↓	G0↓	RMSE↓	G0↓	RMSE↓	G0↓	RMSE↓	G0↓	RMSE↓	G0↓	RMSE↓
BM (ECCV'24)	10.19	17.32	6.20	10.04	34.44	58.57	24.43	32.74	16.16	24.43	21.80	29.42	47.65	55.53	25.25	34.41
PDF (ICML'24)	11.68	17.89	7.52	9.54	30.52	44.01	17.25	28.03	12.24	**19.31**	6.67	9.28	18.55	31.14	12.26	17.58
CRMT (ICLR'24)	10.34	20.03	5.10	8.41	34.52	49.02	10.95	17.40	11.74	20.51	5.15	8.92	15.97	29.63	12.40	19.39
EAU (CVPR'24)	18.77	32.24	12.02	17.56	30.26	52.23	17.65	28.10	21.65	30.07	28.98	34.25	42.12	50.28	35.94	46.32
Free Lunch (CVPR'25)	**9.57**	17.05	4.94	7.81	33.58	62.97	11.90	19.05	23.22	49.01	4.96	7.74	15.19	30.08	13.31	21.23
HOVQ (AAAI'26)	30.23	39.86	18.29	21.37	41.60	71.40	28.88	34.09	76.21	81.80	15.66	23.07	78.66	83.61	46.73	54.88
CPSC (CVPR'26)	15.20	30.72	22.17	30.91	60.46	94.28	14.02	21.76	73.65	84.10	7.81	12.83	77.64	85.64	43.77	53.49
SPD (Ours)	9.86	**13.04**	**4.76**	**7.24**	**22.33**	**42.82**	**9.99**	**15.83**	**10.83**	19.80	**4.91**	**7.44**	**14.48**	**29.43**	**10.89**	**17.55**

Table 12: Quantitative comparison on the DENSE dataset under clean, Gaussian, and Impulse noise settings.

Method	Clean						Gaussian						Impulse					
	10m↓	20m↓	30m↓	Abs.↓	R_{log} ↓	SI_{log} ↓	10m↓	20m↓	30m↓	Abs.↓	R_{log} ↓	SI_{log} ↓	10m↓	20m↓	30m↓	Abs.↓	R_{log} ↓	SI_{log} ↓
ER-F2D (CVPR'24)	0.728	2.120	2.845	0.246	0.395	0.082	1.623	3.772	4.638	0.622	0.824	0.559	1.235	2.363	2.912	0.310	0.382	0.096
PDF (ICML'24)	0.696	1.919	2.621	0.215	0.371	0.081	1.468	3.540	4.453	0.495	0.836	0.534	0.757	2.073	2.786	0.230	0.402	0.092
CRMT (ICLR'24)	1.005	1.863	2.424	0.226	0.308	0.078	1.592	3.520	4.403	0.492	0.765	0.496	0.701	1.752	**2.261**	0.204	0.338	0.085
EAU (CVPR'24)	0.697	1.846	2.568	0.228	0.359	0.091	1.759	3.542	4.588	0.512	0.862	0.522	0.980	2.235	2.977	0.261	0.355	0.100
HAFUNet (MM'25)	0.650	1.944	2.655	0.208	0.381	0.089	1.727	3.481	4.332	0.534	0.763	0.493	0.825	2.201	2.933	0.251	0.334	0.090
HOVQ (AAAI'26)	0.608	1.675	2.279	0.179	0.308	0.073	1.639	3.687	4.569	0.564	0.844	0.480	0.820	1.871	2.425	0.210	0.334	0.090
CPSC (CVPR'26)	0.682	1.741	2.357	0.199	0.326	0.075	1.561	3.482	4.366	0.507	0.826	0.507	1.100	2.299	2.987	0.286	0.438	0.084
SPD (Ours)	**0.581**	**1.582**	**2.203**	**0.176**	**0.299**	**0.070**	**1.450**	**3.385**	**4.266**	**0.451**	**0.745**	**0.465**	**0.631**	**1.750**	2.399	**0.192**	**0.331**	**0.083**

E LIMITATION DISCUSSION

There several limitations left to be studied. Our theoretical analysis uses a common squared error risk and requires an explicit NIG regularization budget, bounded aleatoric variance, uniformly bounded nondegenerate features, a matched homogeneous readout, and a bounded regression domain. These conditions do not automatically follow from NIG validity or a positive soft-

regularization coefficient, and their applicability to the experimental implementations requires separate verification. Finally, SPD still relies on properly learned predictive representations and suitable regularization parameters, which may require further adaptation under extremely noisy or missing-modality conditions.