DYNAMICS-GUIDED DISTRIBUTION SHAPING FOR RELIABLE VISION-LANGUAGE GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language grounding localizes visual regions described by natural language, yet current systems rarely indicate when their localization predictions are unreliable. Existing single-pass probabilistic regression methods mainly estimate uncertainty from the current prediction state and therefore overlook how sample-wise grounding responses evolve during optimization. We propose Dynamics-guided Distribution Shaping (DDS), which uses inter-epoch prediction variation between two consecutive epochs as a local training-dynamics signal. DDS converts this variation into a sample-specific reliability coefficient that guides bounded prediction-distribution shaping and strengthens residual correction for unstable samples. The dynamics branch is used only during training, preserving single-forward-pass localization and uncertainty estimation at inference. Across RefCOCOg and RefCOCO+ with three multimodal backbones, DDS consistently improves uncertainty–error alignment while maintaining competitive grounding performance. Under zero-adaptation transfer from RefCOCO+ to Ref-Adv-s, DDS further improves uncertainty reliability and out-of-distribution discrimination under semantic and reasoning shift.

1 INTRODUCTION

Vision-language grounding (VLG) aims to localize visual content referred to by natural-language expressions, including bounding-box prediction in image grounding and temporal boundary prediction in video grounding Mao et al. (2016); Gao et al. (2017); Hendricks et al. (2017); Deng et al. (2021); Yang et al. (2022); Shi et al. (2024). Recent vision-language models have achieved strong localization performance through improved region–text or moment–text alignment. However, most existing methods mainly optimize localization accuracy, while paying much less attention to whether the model can assess the reliability of its own grounding predictions.

Reliability is crucial for grounding tasks. VLG is inherently ambiguous: multiple plausible regions or moments may satisfy a natural-language expression under different visual contexts, and this ambiguity is further exacerbated by occlusion, small target regions, and unclear temporal boundaries in videos Zhang et al. (2018); Yu et al. (2016); Gao et al. (2017); Hendricks et al. (2017). As a result, a model may produce a plausible localization result with high confidence, even when the prediction is inaccurate or unstable. Standard metrics such as Acc@0.5 and mIoU evaluate final localization quality, but they do not reflect whether the model is aware of uncertain or unreliable cases Guo et al. (2017); Kendall & Gal (2017).



(a) Stable trajectory (b) Unstable trajectory

Figure 1: Representative RefCOCOg validation samples from the deterministic Gemma 4 baseline. Despite similar final IoU (0.947 vs. 0.934), the two samples exhibit substantially different late-stage prediction variation (mean IEPV: 0.009 vs. 0.104). Green denotes the ground truth; dashed cyan and orange denote predictions at epochs 7 and 8; solid blue denotes the epoch-9 prediction.

Existing uncertainty estimation methods include Bayesian approximations, Monte Carlo dropout, deep ensembles, and evidential learning Gal & Ghahramani (2016); Lakshminarayanan et al. (2017);

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.

Amini et al. (2020). Compared with sampling- or ensemble-based methods, evidential regression is more efficient because it predicts uncertainty-related distribution parameters in a single forward pass. This makes it suitable for grounding tasks, where spatial boxes and temporal boundaries are continuous localization outputs. However, the conventional evidential regression primarily estimates uncertainty from the current or converged prediction, making it result-oriented rather than process-aware. It does not account for how localization predictions evolve during training. This distinction is important: two samples may achieve similar final localization errors, yet one may converge smoothly while the other fluctuates across epochs. Such instability indicates unreliable grounding and should be reflected as lower confidence through more conservative evidence. As illustrated in Fig. 1, two samples can attain similarly accurate final boxes while exhibiting much different adjacent-epoch dynamics.

To address this issue, we propose Dynamics-guided Distribution Shaping (DDS) for reliable vision-language grounding. DDS measures the prediction variation of the same sample across two consecutive epochs using Inter-Epoch Prediction Variation (IEPV), and converts it into a sample-wise reliability signal. This signal guides bounded distribution shaping and instability-weighted residual correction, allowing unstable samples to receive more conservative uncertainty estimates while preserving single-pass inference.

We evaluate DDS on RefCOCOg and RefCOCO+ with three multimodal backbones, and further examine zero-adaptation transfer to Ref-Adv-s under semantic and reasoning shift. DDS consistently improves uncertainty–error alignment while maintaining competitive localization performance.

Our main contributions are:

- We propose DDS, which incorporates adjacent-epoch prediction dynamics into reliability-aware vision-language grounding.
- We introduce IEPV as a decay-free sample-wise reliability signal for prediction distribution shaping and residual correction, without additional inference-time passes.
- We demonstrate consistent uncertainty improvements across multiple backbones and datasets, including zero-adaptation evaluation under distribution shift.

2 Related Work

2.1 Vision-Language Grounding

Vision-language grounding (VLG) aims to localize visual content referred to by a natural-language query. In image-based grounding, this problem is commonly studied as visual grounding or referring expression comprehension, where the goal is to predict the bounding box of the target region described by a sentence Yu et al. (2016; 2018); Kamath et al. (2021); Liu et al. (2024); Xiao et al. (2024). Early methods mainly relied on modular attention, proposal-based cross-modal matching, or region-level scoring, while recent approaches increasingly adopt transformer architectures, grounded pre-training, and unified vision-language representations to improve fine-grained region–text alignment. These developments have substantially improved localization accuracy, but most methods still focus primarily on the final grounding prediction rather than on estimating whether that prediction is reliable.

Beyond image-based grounding, VLG has also been extended to videos, where the model is required to localize the temporal segment described by a language query in an untrimmed video Gao et al. (2017); Yuan et al. (2019); Zhang et al. (2020; 2021); Wu et al. (2025). Existing video grounding methods have evolved from proposal-based matching and boundary regression toward stronger cross-modal interaction architectures, including 2D temporal localization, transformer-based alignment, and multimodal large language model based grounding frameworks. Despite their architectural differences, both image-level and video-level grounding involve continuous localization outputs, such as bounding-box coordinates or temporal boundaries. These predictions can be inherently ambiguous because of imprecise language descriptions, complex visual correspondence, occlusion, and uncertain spatial or temporal boundaries. Such ambiguity motivates not only accurate localization, but also reliable estimation of when a grounding prediction should be trusted.

2.2 Uncertainty Estimation and Evidential Regression

Uncertainty estimation in deep learning has been widely studied through Bayesian neural networks, Monte Carlo dropout, deep ensembles, and evidential learning Blundell et al. (2015); Gal & Ghahramani (2016); Lakshminarayanan et al. (2017); Amini et al. (2020). Bayesian neural networks model uncertainty by placing distributions over network weights, Monte Carlo dropout estimates uncertainty through repeated stochastic forward passes, and deep ensembles exploit the diversity of multiple independently trained models. Although effective, these approaches typically require repeated sampling, multiple forward passes, or multiple model instances, which increases computational and memory cost and limits their practicality for large vision-language models.

Evidential deep learning provides a more efficient single-pass alternative by directly predicting the parameters of a higher-order distribution Sensoy et al. (2018); Amini et al. (2020). For regression, a common formulation uses a Normal-Inverse-Gamma distribution parameterized by $\{\gamma, \nu, \alpha, \beta\}$, where γ denotes the prediction mean and the remaining parameters characterize prediction uncertainty. This formulation is naturally compatible with VLG because bounding-box coordinates and temporal boundaries are continuous outputs, allowing localization and uncertainty to be produced jointly in one forward pass.

In this work, we study whether training dynamics can provide an additional reliability cue for vision-language grounding. We adopt evidential regression as our primary single-pass probabilistic instantiation, while the proposed dynamics signal itself depends only on the evolution of localization responses and is therefore conceptually independent of a particular uncertainty head. Standard evidential regression objectives are primarily outcome-oriented: uncertainty is learned from the prediction distribution and localization residual at the current training state, without explicitly representing how an individual prediction evolves during optimization. This distinction is particularly relevant to VLG, where two samples may eventually reach similar localization errors while exhibiting substantially different prediction trajectories across epochs. A sample that fluctuates strongly during training may reflect less stable cross-modal alignment than one that converges smoothly, even when their final localization quality is similar. Motivated by this observation, we use adjacent-epoch prediction variation as a complementary reliability signal so that training dynamics can directly influence the learned prediction distribution.

3 Method

This section presents DDS for reliable vision-language grounding. Figure 2 summarizes its training and inference pipelines. A vision-language backbone and a distribution parameter head first produce the current localization response $\hat{\mathbf{y}}_i^{(t)} = \boldsymbol{\gamma}_i^{(t)}$ together with the prediction-distribution parameters. During training, the IEPV branch compares this response with a sample-indexed, detached previous-epoch prediction $\mathbf{p}_i^{(t-1)}$ and maps their normalized variation $d_i^{(t)}$ to a reliability coefficient $w_{\mathrm{d},i}^{(t)}$. DDS uses this coefficient to control a bounded distribution-shaping factor and an auxiliary residual-correction term. The current detached response then replaces the stored prediction for the next epoch.

Our primary instantiation uses an independent Normal-Inverse-Gamma (NIG) distribution for each localization dimension. For coordinate $j \in \{1, \ldots, D\}$, the head outputs $(\gamma_{ij}, \nu_{ij}, \alpha_{ij}, \beta_{ij})$, where $\nu_{ij} > 0$, $\alpha_{ij} > 1$, and $\beta_{ij} > 0$. The localization prediction is $\hat{y}_{ij} = \gamma_{ij}$, and the coordinate-wise aleatoric and epistemic uncertainties are

$$u_{ij}^{\mathrm{ale}} = \frac{\beta_{ij}}{\alpha_{ij} - 1}, \qquad u_{ij}^{\mathrm{epi}} = \frac{\beta_{ij}}{\nu_{ij}(\alpha_{ij} - 1)}. \tag{1}$$

Unless otherwise stated, we use the mean epistemic uncertainty $u_i^{\mathrm{epi}} = D^{-1} \sum_{j=1}^{D} u_{ij}^{\mathrm{epi}}$ as the scalar uncertainty score. IEPV itself depends only on successive localization responses and is independent of this uncertainty parameterization. At inference, the prediction buffer and all dynamics-guided objectives are removed; a single forward pass produces the localization prediction and the raw NIG parameters used in Eq. (1).

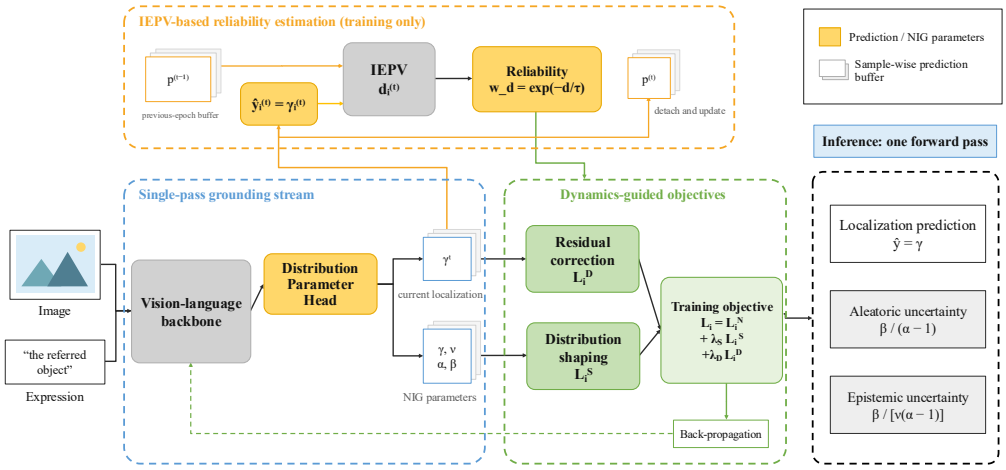


Figure 2: Training and inference framework of Dynamics-guided Distribution Shaping (DDS). During training, DDS compares the current localization response with the detached previous-epoch buffer and converts IEPV into a reliability coefficient that controls bounded distribution shaping and residual correction. The detached current response updates the buffer for the next epoch. At inference, the training-only branch is removed, and one forward pass produces the localization prediction together with aleatoric and epistemic uncertainty.

3.1 Inter-Epoch Prediction Variation

To make prediction uncertainty sensitive to the convergence behavior of grounding predictions, we model the temporal stability of each sample during training. For a sample i, let $\hat{\mathbf{y}}_i^{(t)} \in \mathbb{R}^D$ denotes its localization prediction at epoch t, where $D = 4$ for the bounding-box coordinates considered in our experiments. A stable prediction trajectory indicates that the model has formed a reliable localization pattern for the sample, whereas a highly fluctuating trajectory suggests that its prediction distribution should be more conservative.

For each sample, we store only its detached prediction from the immediately preceding epoch:

$$\mathbf{p}_i^{(t)} = \mathrm{sg}\left(\hat{\mathbf{y}}_i^{(t)}\right), \tag{2}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. The buffer is initialized after the first epoch, and the dynamics terms are enabled from the second epoch onward.

We define IEPV as the normalized deviation from the previous-epoch response:

$$d_i^{(t)} = \frac{\left\| \mathrm{sg}\left(\hat{\mathbf{y}}_i^{(t)}\right) - \mathbf{p}_i^{(t-1)} \right\|_1}{\left\| \mathbf{p}_i^{(t-1)} \right\|_1 + \epsilon}, \tag{3}$$

where ϵ is a small constant for numerical stability, and $\|\cdot\|_1$ denotes the element-wise absolute deviation summed over localization dimensions. A larger $d_i^{(t)}$ indicates that the localization prediction deviates more strongly from its previous trajectory, suggesting the model has not yet formed a stable prediction pattern for this sample.

We convert IEPV into a dynamics reliability coefficient:

$$w_{\mathrm{d},i}^{(t)} = \mathrm{sg}\left[\exp\left(-\frac{d_i^{(t)}}{\tau}\right)\right], \tag{4}$$

where τ is a temperature parameter controlling the sensitivity of the weighting function. The resulting $w_{\mathrm{d},i}^{(t)} \in (0, 1]$ serves as a local dynamics-aware reliability signal. Both the current and previous

responses are treated as constants in this branch, and the outer stop-gradient makes explicit that gradients from the training objective do not flow through IEPV or its reliability coefficient. This prevents the model from reducing its loss by manipulating its current prediction solely to change its own sample weight.

3.2 Bounded Dynamics-Guided Distribution Shaping

Our NIG instantiation characterizes the confidence associated with coordinate j through the concentration score $c_{ij} = 2\nu_{ij} + \alpha_{ij}$. Holding the other parameters fixed, larger ν_{ij} or α_{ij} corresponds to a more concentrated prediction. DDS couples this distributional quantity with sample-wise training dynamics by mapping the reliability coefficient to the bounded shaping factor

$$a_i^{(t)} = 1 + \kappa_{\mathrm{S}} \left(1 - w_{\mathrm{d},i}^{(t)}\right), \qquad \kappa_{\mathrm{S}} \geq 0, \tag{5}$$

where $1 \leq a_i^{(t)} \leq 1 + \kappa_{\mathrm{S}}$. Stable samples have $w_{\mathrm{d},i}^{(t)} \approx 1$ and therefore retain the base shaping strength. As instability increases, $w_{\mathrm{d},i}^{(t)}$ decreases and the shaping strength increases, while remaining bounded by $1 + \kappa_{\mathrm{S}}$. In all experiments, we fix $\kappa_{\mathrm{S}} = 1$ a priori, yielding $a_i^{(t)} \in [1, 2)$ because $w_{\mathrm{d},i}^{(t)} > 0$, without dataset- or backbone-specific tuning.

The dynamics-guided distribution shaping objective is

$$\mathcal{L}_i^{\mathrm{S}} = a_i^{(t)} \sum_{j=1}^{D} |y_{ij} - \gamma_{ij}| \left(2\nu_{ij} + \alpha_{ij}\right). \tag{6}$$

For fixed $a_i^{(t)}$, the gradients of this term with respect to the NIG concentration parameters satisfy

$$\frac{\partial \mathcal{L}_i^{\mathrm{S}}}{\partial \nu_{ij}} = 2a_i^{(t)} |y_{ij} - \gamma_{ij}|, \qquad \frac{\partial \mathcal{L}_i^{\mathrm{S}}}{\partial \alpha_{ij}} = a_i^{(t)} |y_{ij} - \gamma_{ij}|. \tag{7}$$

Thus, an unstable sample receives stronger downward optimization pressure on both ν_{ij} and α_{ij} whenever its localization residual is nonzero, producing a less concentrated prediction distribution. Importantly, Eq. (6) acts directly on the NIG parameters used at inference; the training-time dynamics branch is not required when uncertainty is computed at test time.

3.3 Instability-Weighted Residual Correction

The distribution-shaping objective reduces prediction concentration for unstable samples, but it does not by itself explicitly increase the localization gradient applied to their prediction means. We therefore retain a complementary instability-weighted residual penalty:

$$\mathcal{L}_i^{\mathrm{D}} = \left(1 - w_{\mathrm{d},i}^{(t)}\right) \|\mathbf{y}_i - \boldsymbol{\gamma}_i\|_1. \tag{8}$$

This term assigns larger auxiliary localization penalties to samples with lower temporal consistency. Unlike the distribution-shaping objective, it is independent of the predicted concentration and therefore maintains direct correction of unstable localization responses. Together, the two terms form a dual mechanism: Eq. (6) adjusts prediction concentration, while Eq. (8) strengthens localization correction. The latter targets unstable samples; it does not by itself resolve the distinct stable-but-wrong failure mode.

3.4 Loss Function

The NIG head is trained with the coordinate-wise negative log-likelihood, summed over localization dimensions to obtain $\mathcal{L}_i^{\mathrm{N}}$. On top of this term, we introduce the dynamics-guided distribution shaping objective $\mathcal{L}_i^{\mathrm{S}}$ and the instability-weighted residual term $\mathcal{L}_i^{\mathrm{D}}$. The overall training objective for sample i is defined as

$$\mathcal{L}_i = \mathcal{L}_i^{\mathrm{N}} + \lambda_{\mathrm{S}} \mathcal{L}_i^{\mathrm{S}} + \lambda_{\mathrm{D}} \mathcal{L}_i^{\mathrm{D}}, \tag{9}$$

where λ_{S} and λ_{D} control distribution shaping and residual correction, respectively. During the first epoch, no previous response is available, so we set $w_{\mathrm{d},i}^{(t)} = 1$. Consequently, $a_i^{(t)} = 1$ and $\mathcal{L}_i^{\mathrm{D}} = 0$,

reducing the objective to the base NIG training objective. From the second epoch onward, the dynamics terms are enabled and the detached current response is written to the sample-indexed buffer after each forward pass.

4 EXPERIMENTS

We evaluate DDS on RefCOCOg and RefCOCO+ for in-distribution image grounding and on Ref-Adv-s for zero-adaptation out-of-distribution (OOD) transfer. The experiments assess both localization quality and uncertainty reliability, with particular emphasis on uncertainty–error alignment. We further investigate the robustness of DDS across different multimodal backbones and under distribution shift.

4.1 DATASETS AND METRICS

We use RefCOCOg and RefCOCO+ as the main image-grounding benchmarks Lin et al. (2014); Mao et al. (2016); Yu et al. (2016); RefCOCO+ removes location words and is reported on testA/testB. For zero-adaptation OOD evaluation, models trained on RefCOCO+ are tested directly on the public Ref-Adv-s subset without fine-tuning or threshold selection Dong et al. (2026).

Following the main table, uncertainty quality is measured by Pearson correlation between prediction uncertainty and localization error $(1 - \mathrm{IoU})$ and by AUSE Ilg et al. (2018); higher Pearson and lower AUSE are better. Grounding quality is measured by Acc@0.5 and mIoU. For OOD evaluation, we report the relative accuracy drop, AUROC, and AUPR-OOD, pooling RefCOCO+ testA/testB as ID and treating Ref-Adv-s as the positive OOD class.

4.2 IMPLEMENTATION DETAILS

We use Qwen3.5-9B, Gemma 4 12B, and LFM2.5-VL-3B as backbones Qwen Team (2026); Gemma Team (2026); Liquid AI (2026), adapted with LoRA Hu et al. (2022). We compare deterministic regression, MC Dropout Gal & Ghahramani (2016), deep ensembles Lakshminarayanan et al. (2017), an NIG baseline Amini et al. (2020), and DDS under matched task and adaptation settings. The NIG baseline and DDS share the same prediction head. All three backbones are evaluated on RefCOCOg and RefCOCO+ under the same comparison protocol.

Training is performed for 20 epochs using AdamW with an initial learning rate of 1×10^{-4}, cosine learning rate decay, and a warmup ratio of 0.05. The per-device batch size is set to 1, with gradient accumulation over 16 steps. To improve memory efficiency, we employ mixed-precision training in `bfloat16` together with gradient checkpointing. For the main experiments, we set $\tau = 1.0$, $\kappa_{\mathrm{S}} = 1$, $\lambda_{\mathrm{S}} = 0.3$, and $\lambda_{\mathrm{D}} = 0.3$; the controlled ablation in Table 3 uses $\tau = 1.0$. The sample-wise buffer stores only the detached prediction from the immediately preceding epoch. All experiments are implemented with the Hugging Face Transformers framework Wolf et al. (2020) and conducted on NVIDIA RTX 5090 GPUs.

4.3 COMPARISON WITH OTHER METHODS

Table 1 compares DDS with deterministic regression, MC Dropout, deep ensembles, and the matched NIG baseline across three backbones. DDS achieves the best Pearson and AUSE on every RefCOCOg and RefCOCO+ split. On RefCOCOg, it improves Pearson over the NIG baseline from 0.6751 to 0.7322 for Qwen3.5, from 0.6590 to 0.7147 for Gemma 4, and from 0.5081 to 0.5516 for LFM2.5-VL, while reducing AUSE in all three cases. Unlike MC Dropout and deep ensembles, DDS retains single-pass inference.

The gains persist on RefCOCO+: DDS improves both uncertainty and localization metrics over the NIG baseline on testA/testB for all three backbones. With the compact LFM2.5-VL model, it reaches 0.6613/0.6591 Pearson, 0.0456/0.0552 AUSE, 92.72%/84.31% Acc@0.5, and 81.24%/73.77% mIoU, showing that the dynamics signal is not limited to larger backbones. The only non-best localization entry is RefCOCOg mIoU with LFM2.5-VL (55.01% versus 55.09% for the NIG baseline), so we do not claim uniform localization gains.

Table 1: Main results on RefCOCOg and RefCOCO+ across three backbones. RefCOCO+ entries are testA/testB. Pear.: Pearson correlation; Det.: deterministic regression; MC Drop.: MC Dropout; Ens.: deep ensembles; Evi.: evidential regression. "n/a" denotes an inapplicable metric. ↑ indicates that a higher value is better; ↓ indicates that a lower value is better. **Bold** font denotes the best performance in each category. * denotes localization metrics reported to verify that improved uncertainty estimation does not come at the cost of lower grounding performance; our primary focus is uncertainty quality measured by Pearson and AUSE.

Backbone	Method	RefCOCOg				RefCOCO+			
		Pear.↑	AUSE↓	Acc.* ↑	mIoU* ↑	Pear.↑	AUSE↓	Acc.* ↑	mIoU* ↑
Qwen3.5	Det.	n/a	n/a	91.14	80.15	n/a	n/a	92.74/84.97	82.03/74.73
	MC Drop.	0.3229	0.0845	91.07	79.28	0.3053/0.3231	0.1148/0.1205	93.10/85.35	82.01/74.83
	Ens.	0.5417	0.0403	91.26	80.30	0.4932/0.5294	0.0683/0.0789	93.00/85.25	82.18/74.80
	Evi.	0.6751	0.0173	91.33	80.86	0.6417/0.6841	0.0454/0.0567	92.90/85.12	82.29/74.98
	DDS (Ours)	**0.7322**	**0.0157**	**91.46**	**81.57**	**0.6960/0.7409**	**0.0409/0.0515**	**93.14/85.40**	**82.46/75.14**
Gemma 4	Det.	n/a	n/a	88.43	77.44	n/a	n/a	85.78/80.49	76.16/70.70
	MC Drop.	0.3135	0.1592	88.53	79.50	0.2927/0.2955	0.1611/0.1568	85.98/80.42	76.12/70.61
	Ens.	0.5063	0.0769	88.65	77.32	0.4728/0.4791	0.1146/0.1004	86.02/81.03	77.90/71.88
	Evi.	0.6590	0.0625	89.06	79.26	0.6153/0.6238	0.0959/0.0847	88.29/81.33	78.28/72.57
	DDS (Ours)	**0.7147**	**0.0495**	**89.48**	**80.47**	**0.6674/0.6764**	**0.0872/0.0770**	**88.30/82.08**	**79.70/73.81**
LFM2.5-VL	Det.	n/a	n/a	62.88	54.77	n/a	n/a	92.07/83.21	77.26/69.50
	MC Drop.	0.2419	0.1758	62.95	54.92	0.2900/0.2892	0.0734/0.0850	92.22/83.10	77.20/69.45
	Ens.	0.3908	0.1293	63.40	54.83	0.2682/0.2769	0.0740/0.0862	92.30/83.84	78.65/70.99
	Evi.	0.5081	0.1121	63.24	**55.09**	0.5997/0.6027	0.0502/0.0671	92.66/83.87	79.65/72.06
	DDS (Ours)	**0.5516**	**0.1019**	**64.37**	55.01	**0.6613/0.6591**	**0.0456/0.0552**	**92.72/84.31**	**81.24/73.77**

Table 2: Zero-adaptation OOD evaluation from RefCOCO+ to Ref-Adv-s. ID pools RefCOCO+ testA/testB, and Ref-Adv-s is the positive OOD class. Prediction uncertainty is the OOD score. Det.: deterministic regression; Evi.: matched NIG baseline; "n/a": inapplicable metric.

Backbone	Method	ID Acc.↑	OOD Acc.↑	Rel. Drop↓	OOD Pearson↑	OOD AUSE↓	AUROC↑	AUPR-OOD↑
Qwen3.5	Det.	89.16	42.99	51.78	n/a	n/a	n/a	n/a
	Evi.	89.32	46.23	48.24	0.4309	0.1385	0.7267	0.2596
	DDS (Ours)	**89.58**	**48.17**	**46.23**	**0.4995**	**0.1353**	**0.7520**	**0.2776**
Gemma 4	Det.	83.34	41.42	50.30	n/a	n/a	n/a	n/a
	Evi.	85.08	45.61	46.39	0.4087	0.1496	0.7021	0.2258
	DDS (Ours)	**85.44**	**46.59**	**45.47**	**0.4432**	**0.1168**	**0.7246**	**0.2393**
LFM2.5-VL	Det.	87.99	34.76	60.50	n/a	n/a	n/a	n/a
	Evi.	88.61	38.07	57.04	0.3020	0.1851	0.7146	0.2253
	DDS (Ours)	**88.85**	**39.34**	**55.72**	**0.3304**	**0.1619**	**0.7365**	**0.2379**

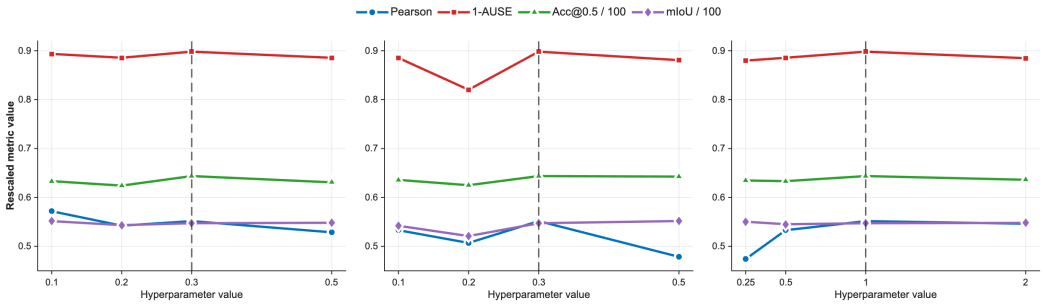
4.4 Performance under Distribution Shift

We evaluate cross-dataset semantic and reasoning shift by training each model on RefCOCO+ and directly testing it on Ref-Adv-s without adaptation. RefCOCO+ testA and testB are pooled only for the ID reference evaluation; no Ref-Adv-s sample is used for training, model selection, threshold tuning, or adaptation. The goal is not only to measure the degradation in grounding accuracy, but also to determine whether prediction uncertainty increases on shifted inputs and can distinguish them from ID samples. This distinction is important because a model may remain temporally consistent while being consistently wrong on an unfamiliar example.

DDS consistently improves over the corresponding NIG baseline under distribution shift. For Qwen3.5, Gemma 4, and LFM2.5-VL, respectively, DDS raises OOD accuracy from 46.23% to 48.17%, 45.61% to 46.59%, and 38.07% to 39.34%, while reducing the relative ID-to-OOD drop. It also improves both uncertainty–error alignment and OOD detection for every backbone, achieving AUROC values of 0.7520, 0.7246, and 0.7365. These results indicate that dynamics-guided shaping improves uncertainty reliability beyond the in-distribution setting without OOD adaptation.

Table 3: IEPV ablation on RefCOCOg with LFM2.5-VL-3B. Both configurations use the same NIG head and training setup. The controlled ablation uses $\tau = 1.0$. * denotes auxiliary localization metrics reported to track grounding performance; Pearson and AUSE are the primary uncertainty metrics.

Dynamics signal	Pear.↑	AUSE↓	Acc.* ↑	mIoU* ↑
	0.3363	0.1296	62.94	54.71
✓	**0.5516**	**0.1019**	**64.37**	**55.01**



(a) λ_S (b) λ_D (c) τ

Figure 3: Hyperparameter sensitivity of DDS on RefCOCOg with LFM2.5-VL-3B. Panels (a)–(c) vary λ_S, λ_D, and τ, respectively, with the other values fixed at $(0.3, 0.3, 1.0)$. Pearson is unchanged; AUSE is shown as $1 - \text{AUSE}$; Acc@0.5 and mIoU are divided by 100. All displayed curves are therefore higher-is-better.

4.5 Ablation Studies

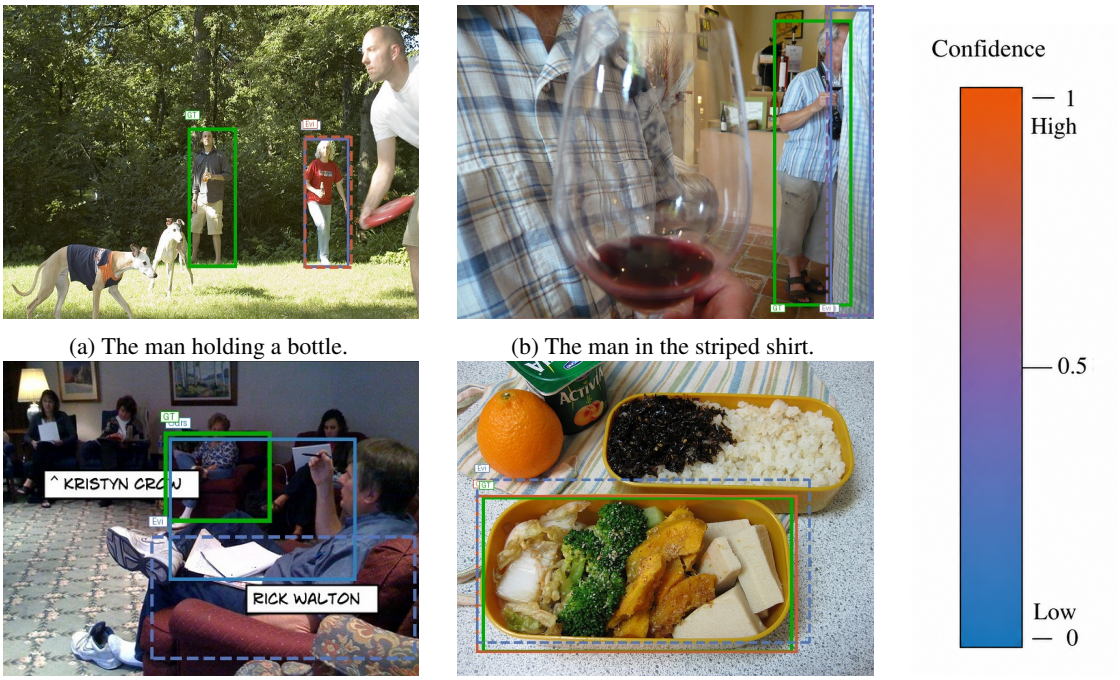
Table 3 isolates the contribution of the complete dynamics branch on RefCOCOg with LFM2.5-VL-3B. Both rows use the same NIG head and training setup; the only change is whether IEPV-guided distribution shaping and residual correction are enabled. Adding DDS increases Pearson from 0.3363 to 0.5516 and reduces AUSE from 0.1296 to 0.1019. It also improves Acc@0.5 from 62.94% to 64.37% and mIoU from 54.71% to 55.01%. Thus, incorporating inter-epoch dynamics improves uncertainty–error alignment and uncertainty ranking without trading off localization quality. Because the two DDS components are activated jointly in this controlled comparison, the table supports the contribution of the complete dynamics branch rather than attributing the gain to either component individually.

Figure 3 evaluates sensitivity to the shaping weight λ_S, residual-correction weight λ_D, and reliability temperature τ. Each sweep varies one value while fixing the other two at $(\lambda_S, \lambda_D, \tau) = (0.3, 0.3, 1.0)$. The selected configuration gives the lowest AUSE and highest Acc@0.5 in the corresponding sweeps, and it maximizes Pearson in the λ_D and τ sweeps. Although $\lambda_S = 0.1$ yields slightly higher Pearson and some alternatives yield marginally higher mIoU, the curves show a consistent trade-off rather than abrupt sensitivity. We therefore select $(0.3, 0.3, 1.0)$ on the validation set and use it for all three backbones without backbone-specific tuning. The bounded shaping coefficient κ_S is fixed to one throughout and is not included in the sweep.

4.6 Uncertainty Analysis

Figure 4 qualitatively compares DDS with the matched NIG baseline on RefCOCOg. In the first three examples, the predicted regions are misaligned with the ground truth, yet the baseline retains relatively high confidence; DDS assigns lower confidence to these unreliable localizations. In the final example, where both predictions are reasonably aligned, DDS preserves an appropriately high confidence rather than uniformly suppressing confidence.

These cases illustrate the intended effect of dynamics-guided shaping: training-time instability is reflected in the prediction distribution while confidence on reliable prediction is retained. Because



(a) The man holding a bottle.

(b) The man in the striped shirt.

(c) The woman labeled KRISTYN CROW.

(d) The lower lunch box.

Figure 4: Qualitative confidence comparison on RefCOCOg. Green boxes denote ground truth; solid and dashed boxes denote predictions from DDS and the matched NIG baseline, respectively. Box color follows the confidence scale shown on the right, with warmer colors indicating higher confidence. In (a)–(c), the baseline remains relatively confident despite localization errors, whereas DDS assigns lower confidence. In (d), both predictions are reasonably aligned and retain appropriate confidence. These selected examples illustrate how confidence changes with localization quality; aggregate uncertainty performance is reported in Table 1.

qualitative examples are necessarily selective, we use the Pearson and AUSE results in Table 1 as the primary evidence for uncertainty reliability.

4.7 Limitations

DDS requires a sample-indexed buffer storing only the previous-epoch prediction for each training sample, resulting in a modest $O(ND)$ training-time storage overhead. Adjacent-epoch variation may also be affected by stochastic optimization or data augmentation, and stable predictions can still be incorrect. Moreover, our distribution-shift evaluation on Ref-Adv-s mainly covers semantic and reasoning shifts rather than broader visual or domain shifts. Finally, although IEPV itself is independent of a particular uncertainty parameterization, our experiments instantiate DDS with a Normal-Inverse-Gamma regression head; extending it to other probabilistic formulations remains future work.

5 Conclusion

We introduced DDS, which uses adjacent-epoch grounding responses to connect local training dynamics with prediction reliability. IEPV provides a decay-free reliability signal that guides prediction distribution shaping and residual correction during training, while inference remains a single forward pass. Across RefCOCOg and RefCOCO+ with three multimodal backbones, DDS consistently improves uncertainty–error alignment while maintaining competitive grounding performance. The gains further extend to zero-adaptation evaluation on Ref-Adv-s, where DDS improves uncertainty reliability and OOD discrimination under semantic and reasoning shift.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and LaTeX formatting. All methodological decisions, experiments, analyses, and scientific claims were independently reviewed and verified by the authors, who take full responsibility for the content of this work.

ETHICS STATEMENT

This work studies uncertainty estimation for vision-language grounding using publicly available benchmark datasets and does not involve human-subject experiments or the collection of personal data. DDS is intended to improve the reliability of model predictions; however, uncertainty estimates should not be interpreted as guarantees of correctness, particularly under distribution shift. We follow the licenses and intended research use of the datasets and pretrained models used in this work.

REPRODUCIBILITY STATEMENT

We provide the model architectures, training objectives, implementation settings, datasets, evaluation metrics, and hyperparameters required to reproduce our experiments in the main paper. The formulation of DDS and all training objectives are described in Section 3, while the experimental protocol and implementation details are provided in Section 4. Code and additional implementation materials will be released with the final version of the paper.

REFERENCES

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14927–14937, 2020.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1613–1622, 2015.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1769–1779, 2021.

Qihua Dong, Kuo Yang, Lin Ju, Handong Zhao, Yitian Zhang, Yizhou Wang, Huimin Zeng, Jianglin Lu, and Yun Fu. Ref-Adv: Exploring MLLM visual reasoning in referring expression tasks. In *International Conference on Learning Representations*, 2026.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, 2016.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5267–5275, 2017.

Gemma Team. Gemma 4 technical report, 2026. URL `https://arxiv.org/abs/2607.02770`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5803–5812, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Eddy Ilg, Özgun Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *Proceedings of the European Conference on Computer Vision*, pp. 652–667, 2018.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR: Modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, pp. 5574–5584, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413, 2017.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Liquid AI. LFM2.5-VL-3B: A better and faster vision-language model for the edge. *Liquid AI Blog*, 2026. https://www.liquid.ai/blog/lfm2-5-vl-3b.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55, 2024.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11–20, 2016.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, volume 31, pp. 3179–3189, 2018.

Fengyuan Shi, Ruopeng Gao, Weilin Huang, and Limin Wang. Dynamic MDETR: A dynamic multimodal transformer decoder for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1181–1198, 2024. doi: 10.1109/TPAMI.2023.3328185.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

Jianlong Wu, Wei Liu, Ye Liu, Meng Liu, Liqiang Nie, Zhouchen Lin, and Chang Wen Chen. A survey on video temporal grounding with multimodal large language model. *arXiv preprint arXiv:2508.10922*, 2025.

Linhui Xiao, Xiaoshan Yang, Fang Peng, Yaowei Wang, and Changsheng Xu. OneRef: Unified one-tower expression grounding and segmentation with mask referring modeling. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4438.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. LAVT: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18155–18165, 2022.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pp. 69–85, 2016.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. MAttNet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1307–1315, 2018.

Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9159–9166, 2019. doi: 10.1609/aaai.v33i01.33019159.

Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4158–4166, 2018.

Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. Multi-stage aggregated transformer network for temporal language localization in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12669–12678, 2021.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2D temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12870–12877, 2020. doi: 10.1609/aaai.v34i07.6984.