ECLAR: EVENT-CONDITIONED VISUAL TOKEN REPAIR FOR LOW-LIGHT VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-light vision-language understanding depends on recovering the visual representations that support correct answers. We introduce ECLAR (Event-Conditioned Latent Alignment and Recovery), which uses synchronized events to repair RGB patch states inside a frozen vision encoder. Four residual adapters align representations with paired clean targets, retaining pretrained readout modules and the visual token count. Inference uses only low-light RGB and events. SDE-VQA-Bench pairs 8,357 reviewed questions with 1,838 frames from 91 sequences to evaluate representation and semantic recovery. With 4.89M parameters trained on SEE-600K, ECLAR leads four final-token recovery metrics against the evaluated restoration cascades, reaching 0.7817 cosine similarity to clean tokens. It improves frozen Qwen3-VL-2B by 11.91 points in six-family Macro and exceeds the strongest cascade by 4.19 points, recovering 57.2% of the clean-to-degraded gap. Matched events add 2.58 points over a capacity-matched RGB-only control; InternVL2.5-2B gains 4.41 points. These results connect clean-token alignment with recovery of both visual representations and downstream semantic performance.

1 INTRODUCTION

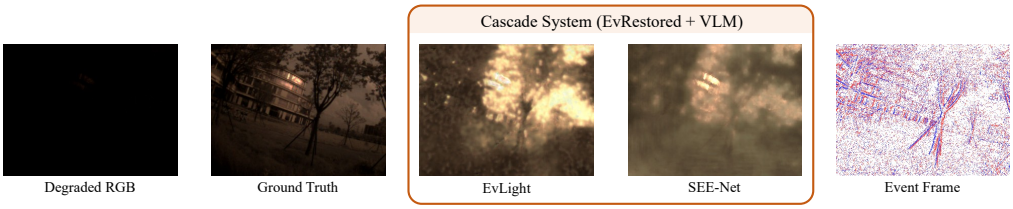
Low light obscures the evidence that vision-language models (VLMs) need to count objects, identify attributes, and describe scenes. Synchronized event cameras provide complementary brightness-change measurements under difficult illumination (Gallego et al., 2022; Liang et al., 2024). Ev-Light and SEE-Net use these measurements to reconstruct RGB images for subsequent interpretation (Liang et al., 2024; Lu et al., 2025). Their image-space objectives target reconstruction quality. For a frozen VLM, the downstream goal is more specific: recover the visual representations that support correct answers.

We formulate this goal as in-encoder visual token repair. Events correct RGB patch states while retaining the pretrained readout modules that connect vision and language (Bai et al., 2025a). Clean-image tokens supply the training targets: recovering these representations aims to restore the visual evidence available to the frozen language model. Event-MLLM aligns fused tokens to clean semantics (Zhang et al., 2026); our design repairs states inside multiple encoder layers (Figure 2).

ECLAR implements this formulation with a shared event encoder and four gated residual adapters inside a frozen vision transformer (ViT). Patch-aligned event evidence conditions each update. The vision encoder, readout modules, and language model remain frozen. Inference uses low-light RGB and events with the original visual token count. Repair locations adapt to each backbone.

SDE-VQA-Bench evaluates whether these representation updates recover semantic performance. It pairs 8,357 reviewed visual question answering (VQA) examples with 1,838 frames from 91 SDE sequences (Liang et al., 2024). RGB references, restoration cascades, and visual token repair answer identical questions about objects, scenes, and text under shared decoding and scoring. ECLAR trains on SEE-600K (Lu et al., 2025), with no SDE question-answer supervision or SDE-based epoch selection.

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.



Prompt: Describe the image in detail.

	Ground Truth	Degraded RGB	Degraded RGB + Event Frame		
Model	Qwen	Qwen	EvLight+Qwen	SEE-Net+Qwen	Ours
Answer	Multi-story building and bare trees	Completely black frame	Microscopic cells	Blurry textured surface	Tree in front of a large-window building

Figure 1: Illustrative SDE scene descriptions under a shared prompt and frozen VLM. Different input-processing paths recover different semantic evidence.

On frozen Qwen3-VL-2B, ECLAR leads four final-token recovery metrics against the evaluated restoration cascades, reaching 0.7817 cosine similarity to clean tokens. It also improves six-family Macro by 11.91 points over Degraded RGB and 4.19 over the strongest cascade. Together, representation and answer-level recovery support clean-token alignment as a training objective. Matched events add 2.58 points over a capacity-matched RGB-only control; InternVL2.5-2B gains 4.41 points.

Our contributions are:

1. **Interface-preserving visual token repair.** Event-conditioned residual adapters repair RGB patch states using paired clean targets while keeping pretrained readout modules and token counts fixed (Section 3).
2. **Paired semantic evaluation.** SDE-VQA-Bench provides a shared question set for comparing sensing and processing paths, with seven question types and explicit annotation provenance (Section 4).
3. **Representation and semantic recovery.** Final-token metrics and VQA comparisons establish recovery at both the visual interface and answer level; controls test event content and repair layouts (Section 5).

2 Related Work

Image enhancement and restoration. Low-light enhancement includes curve estimation (Zero-DCE), Retinex-inspired unrolling (RUAS), and noise-aware processing (Guo et al., 2020; Liu et al., 2021; Xu et al., 2022). RetinexFormer and NAFNet restore RGB images (Cai et al., 2023; Chen et al., 2022a); DA-CLIP conditions restoration on degradation-aware CLIP embeddings (Luo et al., 2024). EvLight and SEE-Net use events for illumination recovery; EFNet uses them for motion deblurring (Liang et al., 2024; Lu et al., 2025; Sun et al., 2022). These methods produce restored images. ECLAR supervises patch states inside a frozen VLM; our cascade comparison measures downstream answers with the same backbone.

Parameter-efficient adaptation. Adapters learn additional modules around frozen pretrained transformations (Houlsby et al., 2019); AdaptFormer applies this strategy to vision transformers (Chen et al., 2022b). Low-rank adaptation (LoRA) learns weight updates, whereas visual prompt tuning learns tokens in the visual input space (Hu et al., 2022; Jia et al., 2022). ECLAR belongs to residual adapter-based adaptation. Its repair operator conditions on synchronized events and RGB states at multiple encoder layers, with paired clean representations as training targets.

Event-aware language models. EventGPT and EventVL learn event-stream language understanding through dedicated event representations and alignment stages (Liu et al., 2025; Li et al., 2025). Reconstruction provides another route to event-based visual question answering (Lou et al., 2025). RE-VLM and LLaFEA combine frame and event encoders, while Talk2Event and EventDrive study grounded understanding and driving (Liu et al., 2026; Zhou & Lee, 2025; Kong et al., 2025; Lu et al., 2026). ECLAR uses events to correct existing RGB tokens inside the vision encoder. The large language model (LLM) receives the native visual sequence, without an additional event-token sequence.

Representation distillation and clean-reference recovery. FitNets transfers intermediate activations, relational knowledge distillation transfers cross-example relations, and contrastive representation distillation uses a contrastive transfer objective (Romero et al., 2015; Park et al., 2019; Tian et al., 2020). Our frozen encoder supplies paired clean targets for degraded-input recovery. Event-MLLM already aligns fused representations to normal-light semantics (Zhang et al., 2026). Building on clean-reference alignment, ECLAR repairs the RGB trajectory inside multiple encoder layers. Qwen3-VL's frozen intermediate DeepStack and final readout modules then process the repaired states (Bai et al., 2025a).

3 Multi-Layer Visual Token Repair

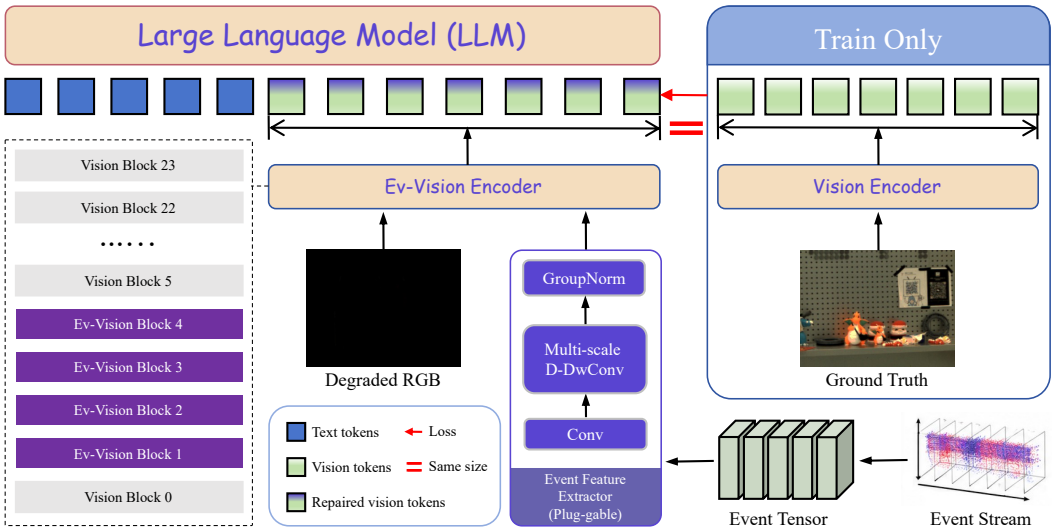


Figure 2: ECLAR with the SDE layout $[1, 2, 3, 4]$. The clean branch shares frozen ViT weights and supplies training targets only. The equality mark denotes matching token shapes. The schematic omits native intermediate readouts, which remain active.

3.1 Repair tokens along the visual trajectory

ECLAR uses events to correct RGB patch states before subsequent frozen layers process them (Figure 2). Let x_d denote a degraded frame, e its synchronized events, and x_c the training-only clean reference. At selected sites $\mathcal{S}$, a trainable residual operator R_θ^ℓ updates the states $Z^\ell \in \mathbb{R}^{P\times d_v}$ from frozen block B_ℓ:

$$Z^\ell = B_\ell(\widehat{Z}^{\ell-1}), \qquad \widehat{Z}^\ell = \begin{cases} Z^\ell + R_\theta^\ell(Z^\ell, e), & \ell \in \mathcal{S}, \\ Z^\ell, & \ell \notin \mathcal{S}. \end{cases} \tag{1}$$

Here $\widehat{Z}^{-1}$ denotes the image's patch embedding. Each native readout M_j maps $P \times d_v$ patch states to language model inputs $H^j = M_j(\widehat{Z}^j) \in \mathbb{R}^{N\times d_l}$. Interface preservation fixes readout weights, locations, token order, and output shapes. Repaired input states change the readout values.

Qwen3-VL-2B uses $d_v = 1024$, $d_l = 2048$, and 2×2 spatial merging (Qwen Team, 2025a). DeepStack reads blocks $\{5, 11, 17\}$; block 23 supplies the final readout (Bai et al., 2025a). Our main sites $\mathcal{S} = \{1, 2, 3, 4\}$ precede the first readout; all indices are zero-based (Table 6).

Qwen2.5-VL-7B repairs follow its four global-attention blocks, $\{7, 15, 23, 31\}$, providing RGB states after cross-window aggregation; the other 28 blocks use windowed attention (Bai et al., 2025b; Qwen Team, 2025b). Table 10 also evaluates early sites $\{1, 2, 3, 4\}$.

3.2 Patch-aligned events drive a bounded residual

We voxelize each event polarity into five temporal bins (ten channels) and align encoded events $u_i \in \mathbb{R}^{256}$ to RGB patches. Statistics $s_i \in \mathbb{R}^3$ contain log positive/negative densities and temporal centroid (Appendix B.1).

Each adapter fuses RGB and event evidence in a 320-dimensional bottleneck (Figure 3). Suppressing the layer index, we compute

$$r_i = W_r \,\mathrm{LN}(z_i), \qquad v_i = m(e)\,[W_e \,\mathrm{RMSNorm}(u_i) + W_s s_i]\,, \tag{2}$$

$$h_i = \mathrm{SiLU}(r_i) \odot \tanh(v_i) + v_i, \qquad g_i = m(e)\,\sigma(q([r_i; v_i; s_i]))\,. \tag{3}$$

LN and RMSNorm denote layer and root-mean-square normalization. Trainable projections W_r, W_e, W_s form the bottleneck; router q and sigmoid σ produce a patch gate. The sample-level mask $m(e)$ indicates a nonempty event stream.

A transport multilayer perceptron (MLP) T maps h_i back to width d_v. The residual update is

$$\widehat{z}_i = z_i + g_i \alpha_\ell T(h_i) \frac{\max(\mathrm{rms}(z_i), 10^{-6})}{\max(\mathrm{rms}(T(h_i)), 10^{-3})}, \qquad \alpha_\ell = \tfrac{1}{2}\sigma(a_\ell), \tag{4}$$

where $\mathrm{rms}(z) = (d_v^{-1} \sum_k z_k^2)^{1/2}$. Each repair layer learns its own scalar a_ℓ and hence α_ℓ, shared across patches; g_i varies by patch. The update bounds residual RMS by half the stabilized input RMS and becomes an exact identity for empty events.

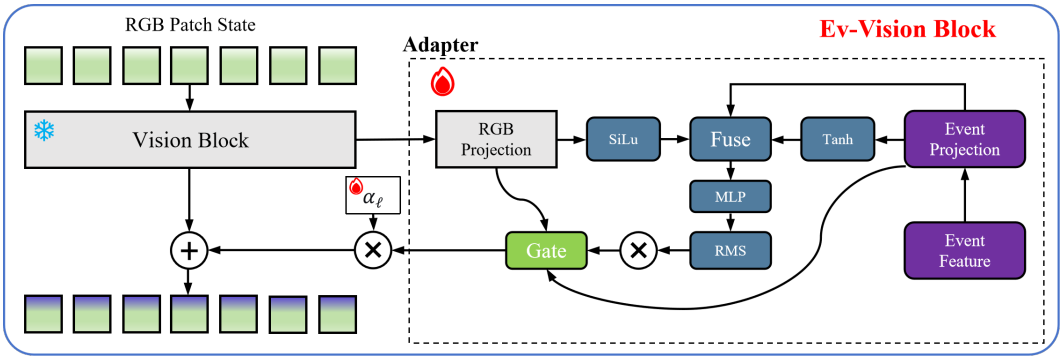


Figure 3: ECLAR block schematic. Squares denote ViT patch states. Each repair layer learns its own α_ℓ, shared across patches; Gate applies patch-dependent g_i. Snowflake/flame denote frozen/trainable components. Appendix B.1 details the schematic conventions.

3.3 Clean states define the recovery target

Frozen clean targets Z_c^ℓ and $H_c = H_c^{23}$ supervise repaired patches and the final output $H = H^{23}$:

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{|\mathcal{S}|} \sum_{\ell \in \mathcal{S}} \mathcal{D}_p(\widehat{Z}^\ell, Z_c^\ell) + \mathcal{D}_f(H, H_c). \tag{5}$$

The patch discrepancy $\mathcal{D}_p$ combines cosine distance, Smooth-L1, and token-relation matching; $\mathcal{D}_f$ also matches final token norms. Teacher targets carry no gradients, and intermediate DeepStack readouts receive no explicit loss.

Event-statistic reconstruction E, mean-gate penalty $\bar{g}$, clean-input identity $\mathcal{L}_{\mathrm{id}}$, and matched-event ranking $\mathcal{L}_{\mathrm{rank}}$ supplement recovery:

$$\mathcal{L}_t = \mathcal{L}_{\mathrm{rec}} + 0.05E + 0.01\bar{g} + \eta_t(\mathcal{L}_{\mathrm{id}} + \mathcal{L}_{\mathrm{rank}}), \tag{6}$$

where $\eta_t = \mathbf{1}[t \bmod 2 = 0]$ activates identity and ranking every second microstep. Appendix B.3 gives the distance definitions, component weights, and auxiliary objectives.

SEE-to-SDE training learns 4.89M parameters from paired SEE-600K representations (Lu et al., 2025), without SDE answer supervision. Inference omits the teacher and auxiliary losses; driving uses separate instruction supervision (Appendix B).

4 SDE-VQA-Bench Measures Semantic Recovery

Paired sensing isolates the input path. SDE-VQA-Bench links clean RGB, low-light RGB, and synchronized events from SDE (Liang et al., 2024) to identical semantic questions. The reviewed release covers 8,357 questions over 1,838 frames from 43 indoor and 48 outdoor sequences (Table 1). All sequences serve evaluation; no SDE example enters ECLAR training. Within each backbone, conditions share prompts, decoding, scoring, and reference answers.

Table 1: Source-domain coverage and question composition of SDE-VQA-Bench.

(a) Coverage by source domain

Domain	Sequences	Frames	Questions
Indoor	43	981	4,363
Outdoor	48	857	3,994
Total	**91**	**1,838**	**8,357**

(b) Composition by question type

Question type	Metric	Questions
Attribute	Acc.	1,269
Counting	Acc.	419
Object Presence	Bal. Acc.	1,522
Object Recognition	Acc.	1,612
Scene Understanding	Acc.	1,654
Text Presence	Bal. Acc.	1,628
OCR Transcription	ANLS	253
Total		**8,357**

Seven tasks measure semantic recovery. Table 1(b) lists seven types and metrics: accuracy (Acc.), balanced accuracy (Bal. Acc.), and average normalized Levenshtein similarity (ANLS) for optical character recognition (OCR). Macro equally weights six families, combining Text Presence and OCR Transcription into one OCR family. Micro averages per-question scores, using ordinary Boolean correctness and normalized edit similarity for text. Appendix C distinguishes these estimators from frequency-weighted balanced accuracy.

Clean references anchor annotation. Qwen3-VL-8B-Instruct (Qwen Team, 2025c) proposes structured facts from clean images without access to Degraded RGB, events, or evaluated-model predictions. Schema checks, deduplication, an independent DeepSeek-V4 Flash Vision Exp audit, and human adjudication of model-label disagreements produce the 8,357-question release (Figure 4). Appendix A details inclusion decisions and the model-assisted annotation scope.

Complete pairing supports direct comparisons. Every main condition and layout covers all 8,357 identifiers, counting invalid answers as errors. The supplementary material includes test annotations and frame-to-sequence mappings. We will publicly release SDE-VQA-Bench and its evaluation toolkit upon acceptance; original SDE assets remain subject to their source terms.

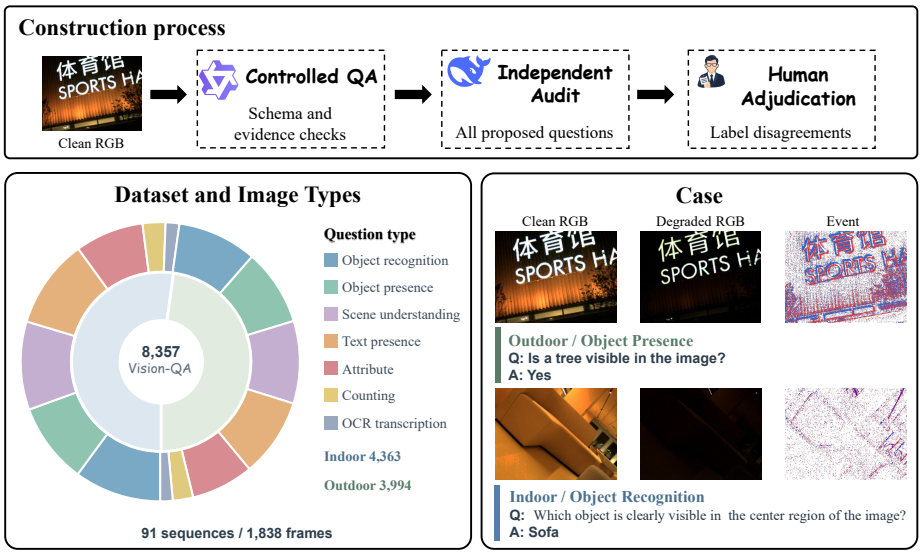


Figure 4: SDE-VQA-Bench overview. Top: clean-image question generation, independent audit, and human adjudication of disagreements. Bottom left: 8,357 questions by setting and question type. Bottom right: clean RGB, degraded RGB, and event inputs paired with example questions and answers.

5 Visual Token Repair Improves Understanding

5.1 Token repair improves aggregate semantic recovery

Where used, **bold** and underlining mark the best and second-best table scores within each comparison group, excluding oracle rows.

Paired protocol. We compare input-processing paths within frozen Qwen3-VL-2B-Instruct and InternVL2.5-2B (Bai et al., 2025a; Chen et al., 2024), sharing questions, decoding, and scoring. EvLight and SEE-Net use RGB and events (Liang et al., 2024; Lu et al., 2025); RetinexFormer and DA-CLIP use RGB alone (Cai et al., 2023; Luo et al., 2024). Their restored images feed the frozen VLM. Qwen3 includes official and SEE-fine-tuned DA-CLIP weights (Appendix B.6) and SEE alignment-stage Event-MLLM retraining (Appendix B.5). ECLAR and the RGB-only control use final three-epoch SEE checkpoints. Appendix F specifies InternVL preprocessing and normalization. Panel (c) reports two-stage Event-MLLM reproduction with Qwen2.5-VL-3B and SDE-source training exposure (Appendix H).

Token repair recovers object evidence. ECLAR gains 11.91 Macro and 9.81 Micro points over Degraded RGB in Qwen3 (Table 2(a)). It exceeds the strongest restoration cascade by 4.19 Macro points (official DA-CLIP) and 3.41 Micro points (DA-CLIP +SEE). Repair recovers 57.2% of the clean-to-degraded Macro gap; the paired 91-sequence bootstrap gives a 95% interval of $[9.29, 14.42]$ points for the Macro gain over Degraded RGB (Appendix C).

Counting rises from 13.60 to 40.57 and object presence from 54.66 to 82.41, locating recovery in object-level answers (Appendix D).

Recovery extends to InternVL2.5-2B. Separately trained ECLAR reaches 58.56 Macro and 59.23 Micro (Table 2(b)), gaining 4.41 Macro over Degraded RGB and 3.53 over RetinexFormer, the strongest evaluated cascade. Appendix F specifies preprocessing and uncertainty.

Repair reaches the final visual tokens. We compare final post-merger tokens against clean RGB references on SDE's 1,838 frames, weighting frames equally. Cosine similarity, relative L2, linear

Table 2: SDE-VQA-Bench (8,357 questions). R/E: RGB/events; Rec: recognition; OCR: mean of text presence and transcription. Ranks use unrounded scores within each backbone. †: SEE alignment-stage retraining. ‡: official-data two-stage reproduction, with SDE-source exposure (Appendix H).

(a) Qwen3-VL-2B-Instruct									
Condition / method	Mod.	Attr.	Count	Pres.	Rec	Scene	OCR	Macro	Micro
Reference inputs									
Clean RGB (oracle)	R	84.16	44.63	88.92	81.95	85.25	69.64	75.76	83.36
Degraded RGB	R	72.50	13.60	54.66	**74.69**	64.75	49.47	54.95	65.29
Cascaded systems: RGB restoration → frozen VLM									
EvLight	R+E	67.38	13.13	66.26	64.21	52.30	57.17	53.41	63.42
RetinexFormer	R	68.72	22.91	72.33	71.34	67.05	**61.40**	60.63	70.12
SEE-Net	R+E	72.58	28.64	75.65	71.65	64.39	60.92	62.30	71.17
DA-CLIP (official)	R	70.76	29.59	78.38	70.10	67.23	59.94	62.67	71.55
DA-CLIP (+SEE)	R	**77.31**	23.63	73.27	71.22	70.44	56.28	62.02	71.69
Final-token fusion									
Event-MLLM†	R+E	74.15	23.63	68.03	66.44	57.50	29.82	53.26	64.14
In-encoder visual token repair									
RGB-only control	R	75.26	33.41	79.09	70.97	71.16	55.77	64.28	73.23
ECLAR $[1, 2, 3, 4]$	R+E	75.97	**40.57**	**82.41**	74.38	**71.46**	56.34	**66.86**	**75.10**
(b) InternVL2.5-2B									
Condition / method	Mod.	Attr.	Count	Pres.	Rec	Scene	OCR	Macro	Micro
Reference inputs									
Clean RGB (oracle)	R	78.88	45.35	79.84	80.96	81.26	31.65	66.32	65.99
Degraded RGB	R	**71.32**	19.81	62.42	69.23	68.38	33.75	54.15	56.94
Cascaded systems: RGB restoration → frozen VLM									
EvLight	R+E	62.33	25.78	63.43	59.99	49.76	30.43	48.62	48.99
RetinexFormer	R	63.12	32.22	68.02	67.87	63.24	**35.69**	55.03	56.94
SEE-Net	R+E	65.72	34.84	65.83	69.04	58.28	29.18	53.82	54.38
In-encoder visual token repair									
ECLAR $[1, 2, 3, 4]$	R+E	71.00	**35.80**	**73.37**	**72.83**	**69.17**	29.19	**58.56**	**59.23**
(c) Qwen2.5-VL-3B: official-data reproduction									
Event-MLLM‡	R+E	60.36	10.50	64.28	60.05	59.13	40.29	49.10	61.27

centered kernel alignment (CKA) (Kornblith et al., 2019), and relation error measure recovery (Appendix J). ECLAR leads all four metrics against the evaluated cascades (Table 3). Together with its VQA gains, this supports clean-token alignment: encoder-level repair recovers the visual representations supplied to the frozen language model and improves aggregate semantic understanding.

The official-data Event-MLLM reproduction reaches 49.10 Macro and 61.27 Micro; Appendix H validates original-task performance and specifies backbone and training-exposure differences.

5.2 Event evidence supports repair across layouts

Event content adds recovery beyond capacity. Matched events add 2.58 Macro and 1.87 Micro points over the Qwen3 RGB-only control (Table 4), which replaces event voxels with ones during training and inference. Both use 4.89M parameters at $[1, 2, 3, 4]$, with matched backbone, data, losses, and nominal budget (Appendix B). Events improve all six families, led by Counting (+7.16).

Readout-aware controls test repair locations. Three layouts target Qwen3's DeepStack/final readouts or preceding blocks (Bai et al., 2025a; Qwen Team, 2025a); three test early sites (Table 6). All share architecture, frozen backbone, SEE data, losses, and nominal budget; GPU/resume schedules differ (Appendix B). We evaluate final checkpoints without SDE training or epoch selection.

Table 3: Visual token representation recovery on the 1,838 SDE-VQA-Bench frames. Each frame contributes equally. Metrics compare final post-merger tokens with the corresponding clean RGB tokens in Qwen3-VL-2B. Scores use their native scales. Clean RGB provides the reference; ranks exclude this row.

Method	Cosine ↑	Relative L2 ↓	Linear CKA ↑	Relation error ↓
Clean RGB (reference)	1.0000	0.0000	1.0000	0.0000
Degraded RGB	0.7232	0.8582	0.3352	0.1432
RetinexFormer	0.7079	0.9613	0.3198	0.1396
EvLight	0.6884	0.9695	0.2929	0.1439
SEE-Net	0.7343	0.9100	0.3394	0.1360
DA-CLIP (official)	0.7197	0.8869	0.3189	0.1437
DA-CLIP (+SEE)	0.7515	0.8379	0.3621	0.1338
ECLAR	**0.7817**	**0.6884**	**0.4829**	**0.1203**

Table 4: Event-input control at $[1, 2, 3, 4]$. Repair variants train 4.89M parameters.

Variant	Input	Macro	Micro
Degraded RGB	None	54.95	65.29
RGB-only control	Ones	64.28	73.23
ECLAR	Matched	**66.86**	**75.10**

Table 5: Event-encoder substitution at $[1, 2, 3, 4]$. M: total trainable parameters.

Encoder	M	Macro	Micro
Lightweight CNN	4.89	**66.86**	**75.10**
ResNet-18	15.95	63.73	72.79
EFNet-derived	6.38	60.87	70.18
RVT-T (adapted)	7.88	62.02	71.81
Spikformer (adapted)	7.25	63.86	72.81

Table 6: Qwen3-VL-2B injection-placement ablation (four adapters, 4.89M trainable parameters). DeepStack readouts: blocks 5, 11, 17; final readout: block 23. Indices are zero-based.

Placement	Injection after blocks	Macro	Micro
Readout-adjacent controls			
At DeepStack + final readout blocks	$[5, 11, 17, 23]$	65.29	74.25
One block before all readouts	$[4, 10, 16, 22]$	65.71	**75.12**
Before DeepStack + at final block	$[4, 10, 16, 23]$	65.60	74.44
Early layouts			
Early, ending at first DeepStack block	$[1, 2, 3, 5]$	66.42	75.08
Early, before first DeepStack block (main)	$[1, 2, 3, 4]$	**66.86**	75.10
Early, shifted to block 0	$[0, 1, 2, 3]$	66.72	75.01

Recovery holds across six placements. All six layouts improve Degraded RGB by 10.34–11.91 Macro points (Table 6). We use $[1, 2, 3, 4]$ as the main layout based on its leading SDE Macro; $[4, 10, 16, 22]$ leads Micro. The 1.57-point Macro range shows recovery across early and readout-adjacent sites under the shared clean-state training recipe.

A lightweight encoder suffices. Four alternatives to our convolutional neural network (CNN) gain 5.93–8.91 Macro points over Degraded RGB (Table 5); our lightweight encoder leads Macro with the fewest trainable parameters. These substitutions share the event window, repair sites, and nominal training recipe (Appendix I), supporting compatibility across the tested event encoders.

5.3 Event-conditioned adaptation supports driving

Protocol. Qwen2.5-VL-7B (Bai et al., 2025b) trains 5.13M adapter parameters, optionally adding language LoRA (Hu et al., 2022), at $[7, 15, 23, 31]$ (Appendices B.4 and E). Baselines follow Lu et al. (2026), arXiv v1, Tables 2 and 10; * marks frame-only tuning. Scores are percentages except L2

(meters). QA denotes question answering; grounding uses intersection-over-union (IoU) accuracy at 0.6 and mean IoU.

Table 7: EventDrive comparison. Adapter: adapter-only; +LoRA: joint adaptation. E/R/R+E: events/RGB/both. Bold/underline mark best/second-best displayed scores within each split.

Method	Mod.	Perception			Understanding			Prediction		Planning		
		DSEC	M3ED	PKU	QA	Acc@0.6	mIoU	Speed	Path	Speed	Path	L2↓
(a) Regular split												
EventGPT-7B	E	51.40	54.97	52.25	38.78	5.49	8.24	27.84	65.44	28.99	76.08	11.42
EventDrive-VLM	E	62.51	67.07	62.39	54.21	43.43	47.82	34.68	82.25	42.56	84.64	6.89
LLaVA-v1.6-Mistral-7B	R	58.65	59.24	50.05	40.37	15.02	29.00	18.22	29.68	9.89	87.51	8.13
LLaVA-OneVision-1.5-8B	R	80.81	86.35	72.42	61.62	2.91	12.50	14.38	83.91	56.51	58.50	9.19
InternVL2.5-8B	R	78.42	83.41	75.86	60.02	0.29	2.50	39.72	86.41	12.52	73.34	10.19
InternVL3-8B	R	78.61	86.27	74.37	60.60	0.24	2.31	4.41	91.25	52.04	84.34	9.84
Qwen2.5-VL-7B	R	77.71	76.92	62.46	49.98	63.35	60.11	8.47	81.64	39.32	87.41	8.78
Qwen2.5-VL-7B*	R	81.52	84.69	75.88	58.44	69.94	67.19	36.84	84.94	53.87	89.44	4.54
EventDrive-VLM	R+E	85.44	86.64	78.89	65.46	72.86	**72.56**	42.44	87.49	57.03	92.35	3.66
ECLAR (Adapter)	R+E	92.56	94.50	91.77	57.12	5.82	22.50	44.91	**92.17**	**73.05**	87.39	7.2022
ECLAR + LoRA	R+E	**94.47**	**98.16**	**96.72**	**72.75**	**75.67**	66.40	**45.20**	89.25	55.14	**93.67**	**1.6971**
(b) Hard split												
EventGPT-7B	E	49.09	55.00	52.83	36.96	2.89	3.67	14.26	59.24	24.82	69.64	12.24
EventDrive-VLM	E	67.30	66.22	66.93	52.17	34.22	29.50	28.56	76.48	39.23	85.53	7.26
LLaVA-v1.6-Mistral-7B	R	55.64	61.55	50.45	36.69	15.23	25.30	20.10	23.25	10.43	88.59	7.31
LLaVA-OneVision-1.5-8B	R	80.00	85.36	76.26	57.09	2.65	16.10	15.23	55.84	46.44	41.12	8.31
InternVL2.5-8B	R	75.95	86.27	80.15	58.28	0.18	2.21	27.05	78.52	34.02	51.75	11.08
InternVL3-8B	R	74.11	84.09	75.61	58.68	0.16	1.51	13.71	78.58	33.78	83.62	8.97
Qwen2.5-VL-7B	R	72.25	77.14	60.76	43.05	27.15	35.12	11.68	70.15	43.23	82.62	7.91
Qwen2.5-VL-7B*	R	77.94	83.52	68.88	50.77	38.15	40.89	20.76	75.46	46.02	86.44	6.02
EventDrive-VLM	R+E	82.43	86.09	76.94	59.04	**45.86**	**49.66**	29.48	79.32	52.45	88.64	5.41
ECLAR (Adapter)	R+E	94.36	94.31	93.43	61.46	4.64	23.90	51.27	81.22	**76.14**	90.50	6.5968
ECLAR + LoRA	R+E	**95.38**	**97.99**	**96.92**	**76.82**	42.38	48.20	**52.39**	**82.44**	64.30	**95.52**	**1.4547**

Joint adaptation extends spatial performance. Adapter-only repair exceeds the listed baselines on perception, speed prediction, and speed planning in both splits. Joint Adapter+LoRA training reaches QA accuracies of 72.75/76.82 and trajectory L2 of 1.6971/1.4547 meters on Regular/Hard. Grounding Acc@0.6 reaches 75.67/42.38, compared with 5.82/4.64 for adapter-only training. The joint setting supports spatial outputs alongside categorical decisions; the adapter-only setting retains a frozen VLM. Appendix G probes a separate, earlier adapter-only checkpoint.

6 Scope and Limitations

On EventDrive, adapter-only training leaves a gap in fine-grained grounding and trajectory regression while keeping the VLM frozen. Joint Adapter+LoRA training improves grounding accuracy and reduces trajectory L2 error on both Regular and Hard splits (Table 7), mitigating this limitation through additional language-model adaptation.

7 Conclusion

ECLAR uses events and clean-token supervision to repair visual representations while retaining a frozen VLM's readout modules and token counts. Final-token metrics and paired VQA results support representation and semantic recovery beyond the evaluated restoration cascades. Event-input and placement controls support the use of event evidence across repair layouts.

REPRODUCIBILITY STATEMENT

Appendix B specifies the adapter, loss weights, optimization budget, checkpoint rule, and decoding settings. Appendix C defines task aggregation and the paired sequence bootstrap. Every SDE condition covers all 8,357 reviewed questions. The six placement runs share model capacity and the nominal training budget; Appendix B records their execution differences. Supplementary material includes the anonymized implementation, fixed SDE-VQA-Bench test annotations, frame-to-sequence mappings, and EventDrive training and evaluation code. Its documentation specifies external datasets, model weights, and path configuration. We will publicly release SDE-VQA-Bench and the code upon acceptance.

ETHICS STATEMENT

This work uses existing event-camera research datasets and does not collect personal information. Event-enhanced perception may help systems operating under difficult illumination, but it may also increase sensing capability in surveillance settings. The benchmark labels are partly model generated: human adjudication covers model-label disagreements, not every retained question. This provenance may encode model-family, language, task-template, and scene-distribution biases. Any release will preserve source licenses, provenance, exclusion decisions, and known label-distribution warnings. Safety-critical deployment requires task-specific validation for text recognition, spatial grounding, and trajectory prediction.

AI USE STATEMENT

Generative AI tools assisted literature discovery, reference organization, research-code development, and manuscript drafting and editing. Qwen3-VL-8B-Instruct (Qwen Team, 2025c) proposed benchmark annotations from clean images, and the DeepSeek-V4 Flash Vision Exp vision API audited the proposals. Section 4 and Appendix A specify model involvement and the scope of human adjudication. Deterministic programs enforced the schema and recorded review decisions. The authors take responsibility for the final manuscript, benchmark, and experimental code.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12470–12479, 2023. doi: 10.1109/ICCV51070.2023.01149.

Kenneth Chaney, Fernando Cladera, Ziyun Wang, Anthony Bisulco, M. Ani Hsieh, Christopher Korpela, Vijay Kumar, Camillo J. Taylor, and Kostas Daniilidis. M3ED: Multi-robot, multi-sensor, multi-environment event dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4016–4023, 2023. doi: 10.1109/CVPRW59228.2023.00419.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European Conference on Computer Vision*, volume 13667 of *Lecture Notes in Computer Science*, pp. 17–33, 2022a. doi: 10.1007/978-3-031-20071-7_2.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16664–16678, 2022b. doi: 10.52202/068431-1212.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022. doi: 10.1109/TPAMI.2020.3008413.

Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629(8014):1034–1040, 2024. doi: 10.1038/s41586-024-07409-w.

Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13884–13893, 2023. doi: 10.1109/CVPR52729.2023.01334.

Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. doi: 10.1109/LRA.2021.3068942.

Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1780–1789, 2020. doi: 10.1109/CVPR42600.2020.00185.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799, 2019.

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.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, volume 13693 of *Lecture Notes in Computer Science*, pp. 709–727, 2022. doi: 10.1007/978-3-031-19827-4_41.

Lingdong Kong, Dongyue Lu, Alan Liang, Rong Li, Yuhao Dong, Tianshuai Hu, Lai Xing Ng, Wei Tsang Ooi, and Benoit R. Cottereau. Talk2Event: Grounded understanding of dynamic scenes from event cameras. In *Advances in Neural Information Processing Systems*, volume 38, pp. 38802–38851, 2025. doi: 10.52202/085713-1159.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Pengteng Li, Yunfan Lu, Pinghao Song, Wuyang Li, Huizai Yao, and Hui Xiong. EventVL: Understand event streams via multimodal large language model. *arXiv preprint arXiv:2501.13707*, 2025.

Guoqiang Liang, Kanghao Chen, Hangyu Li, Yunfan Lu, and Lin Wang. Towards robust event-guided low-light image enhancement: A large-scale real-world event-image dataset and novel approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23–33, 2024. doi: 10.1109/CVPR52733.2024.00011.

Hanqing Liu, Mingjie Liu, Luoping Cui, Endian Lin, Donghong Jiang, and Chuang Zhu. RE-VLM: Event-augmented vision-language model for scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10377–10386, 2026.

Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10561–10570, 2021. doi: 10.1109/CVPR46437.2021.01042.

Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Xin Meng, Fei Richard Yu, Xiangyang Ji, and Ming Li. EventGPT: Event stream understanding with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29139–29149, 2025. doi: 10.1109/CVPR52734.2025.02713.

Hanyue Lou, Jiayi Zhou, Yang Zhang, Boyu Li, Yi Wang, Guangnan Ye, and Boxin Shi. Reconstruction as a bridge for event-based visual question answering. *arXiv preprint arXiv:2512.11510*, 2025.

Dongyue Lu, Rong Li, Ao Liang, Lingdong Kong, Wei Yin, Lai Xing Ng, Benoit R. Cottereau, Camille Simon Chane, and Wei Tsang Ooi. EventDrive: Event cameras for vision-language driving intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22312–22322, 2026.

Yunfan Lu, Xiaogang Xu, Hao Lu, Yanlin Qian, Pengteng Li, Huizai Yao, Bin Yang, Junyi Li, Qianyi Cai, Weiyu Guo, and Hui Xiong. SEE: See everything every time – adaptive brightness adjustment for broad light range images via events. *arXiv preprint arXiv:2502.21120*, 2025.

Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B. Schön. Controlling vision-language models for multi-task image restoration. In *International Conference on Learning Representations*, pp. 16226–16246, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019. doi: 10.1109/CVPR.2019.00409.

Qwen Team. Qwen3-VL-2B-Instruct: Model configuration, 2025a.

Qwen Team. Qwen2.5-VL-7B-Instruct: Model configuration, 2025b.

Qwen Team. Qwen3-VL-8B-Instruct: Model card, 2025c.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *European Conference on Computer Vision*, volume 13678 of *Lecture Notes in Computer Science*, pp. 412–428, 2022. doi: 10.1007/978-3-031-19797-0_24.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. SNR-aware low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17714–17724, 2022. doi: 10.1109/CVPR52688.2022.01719.

Baoheng Zhang, Jiahui Liu, Gui Zhao, Weizhou Zhang, Yixuan Ma, Jun Jiang, Yingxian Chen, Wilton W. T. Fok, Xiaojuan Qi, and Hayden Kwok-Hay So. Learning to see through illumination extremes with event streaming in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26198–26208, 2026.

Hanyu Zhou and Gim Hee Lee. LLaFEA: Frame-event complementary fusion for fine-grained spatiotemporal understanding in LMMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22294–22304, 2025. doi: 10.1109/ICCV51701.2025.02070.

Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng Yan, Yonghong Tian, and Li Yuan. Spikformer: When spiking neural network meets transformer. In *International Conference on Learning Representations*, 2023.

A Benchmark Construction Details

We sample every second frame, remove near duplicates using a clean-image difference-hash threshold of six, verify event files, and standardize RGB and event orientation. Deterministic validation checks source integrity, answerability constraints, and a maximum of six questions per frame before strict JSON serialization. Historical source-split labels remain provenance fields, not training assignments for this study.

SDE-VQA-Bench contains 8,357 questions over 1,838 frames from all 91 SDE sequences and passes schema validation with no reported errors. Inclusion requires auditor agreement or human adjudication supporting the original label. We exclude audit-corrected, ambiguous, unsure, and unresolved items. Human adjudication covers 795 model-label disagreements during construction, rather than every retained question. We describe SDE-VQA-Bench as semi-automatically generated, independently audited, and manually adjudicated on model-label disagreements.

Each provenance record includes the source sequence and frame, paired image paths, synchronized events, proposer output, deterministic validation decisions, audit flag, and human adjudication when available. The release retains 6,416 Easy, 1,688 Medium, and 253 Hard questions. These labels describe question construction rather than empirical model difficulty. Answer and scene-category concentration warnings accompany the release to support interpretation of absolute scores.

B Training and Evaluation Details

B.1 Architecture and optimization

Figure 3 depicts one frozen vision block and its event-conditioned residual adapter. Fuse computes $\mathrm{SiLU}(r_i) \odot \tanh(v_i) + v_i$. RMS normalizes the transport output, and the following multiplication scales it by the stabilized RGB patch RMS from Equation 4. The schematic omits this RGB-magnitude input line. Gate computes and applies g_i. The projections and gate incorporate the normalization, event statistics, and presence mask from Equations 2 and 3, which the drawing omits for clarity. Each repair layer has an independent learned a_ℓ, giving $\alpha_\ell = \frac{1}{2}\sigma(a_\ell)$; this scalar scales all patches at that layer.

The Qwen3-VL-2B implementation uses four independent adapters and one shared event encoder. A 5×5 stride-two convolution maps ten event channels to width 256. Three residual blocks apply 3×3 depthwise convolutions with dilation 1, 2, and 4, pointwise projections, GroupNorm, and SiLU. Resampling maps event evidence and statistics to the ViT patch layout. Each repair adapter uses a width-320 bottleneck, one residual transport MLP, one scalar router, and an auxiliary three-dimensional event-statistic decoder. The evaluated model has no second pooled-context residual.

The statistics s_i contain log positive density, log negative density, and temporal centroid. Resampling and reordering align the event map and statistics with the RGB patch order. The sample-level presence mask broadcasts to all patches, allowing convolutional context to support locally sparse

events. Initial router bias -3 and scale parameter $a_\ell = -2$ start the update near the pretrained RGB path. An empty stream sets both v_i and g_i to zero, giving the exact identity in Equation 4.

Training uses SEE-600K, AdamW with learning rate 2×10^{-4} and weight decay 0.05, 500 warmup updates followed by cosine decay, and gradient clipping at norm 1. We use bfloat16, effective batch size eight, three epochs, and seed 3407. The image processor's training pixel limit is 401,408. The backbone remains frozen; the optimizer contains 4.89M event-encoder and repair-block parameters, including auxiliary decoders. Each run contributes its final checkpoint after the fixed training budget, without SDE-based epoch selection.

The three readout-adjacent runs and $[1, 2, 3, 5]$ use one GPU with gradient accumulation eight. The $[1, 2, 3, 4]$ run starts on two GPUs with accumulation four and resumes on four GPUs with accumulation two. The $[0, 1, 2, 3]$ run uses four GPUs with accumulation two. Resumes restore optimizer and scheduler state. The effective batch and training budget agree across runs, but per-rank sample grouping and floating-point reductions can differ. The six results describe single-seed placement behavior, not replicated optimization estimates.

The RGB-only control keeps the $[1, 2, 3, 4]$ architecture and nominal training settings. Before event encoding, it replaces every voxel tensor with an all-ones tensor of the same shape, including the identity and ranking branches. This preserves model capacity and permits nonzero corrections while removing sample-specific event information. Both training and inference use this constant input. The run starts on one GPU with accumulation eight and resumes from update 6,000 on two GPUs with accumulation four; evaluation uses the final checkpoint.

B.2 Parameter accounting across backbones

Parameter counts cover one shared event encoder and four independent repair blocks, including their auxiliary decoders; they exclude the frozen VLM. With ten event channels, event width 256, and bottleneck width 320, the encoder has 271,616 parameters and each repair block has $642d_v + 497,931$. Qwen3-VL-2B uses $d_v = 1024$, giving $271,616+4\times1,155,339 = 4,892,972$ parameters (4.89M). Qwen2.5-VL-7B uses $d_v = 1280$, giving $271,616 + 4 \times 1,319,691 = 5,550,380$ (5.55M). The wider RGB projections and input LayerNorm account for the 657,408-parameter difference; changing the four injection indices alone does not change capacity.

Each auxiliary decoder contains 104,323 parameters, or 417,292 across four sites. These heads predict event statistics for training and diagnostics; their outputs do not feed repaired tokens or language answers. SEE training includes these heads in its 4.89M trainable parameters. Pure supervised fine-tuning (SFT) for driving freezes them, leaving $5,550,380 - 417,292 = 5,133,088$ trainable adapter parameters (5.13M). Joint driving adaptation adds 80,740,352 language LoRA parameters, for 85,873,440 trainable parameters (85.87M). Freezing the auxiliary heads leaves the inference architecture unchanged.

B.3 Complete representation objective

Section 3.3 gives the recovery and total training objectives. Equation 5 averages patch losses over the four repair sites and adds the final-merger loss once. The shared frozen teacher supplies detached clean targets Z_c^ℓ and $H_c = H_c^{23}$; intermediate DeepStack outputs receive no explicit loss.

For token arrays X, Y with aligned positions, let $C(X, Y)$ be mean token cosine distance and $S(X, Y)$ mean Smooth-L1 loss with transition parameter one. Let $R(X, Y)$ be mean absolute error between Gram matrices of L2-normalized tokens, using at most 128 evenly spaced token indices. With n aligned tokens, define

$$N(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\|X_i\|_2}{\max(\|Y_i\|_2, 10^{-6})} - 1 \right|. \tag{7}$$

The discrepancies in Equation 5 are

$$\begin{aligned} \mathcal{D}_p(X, Y) &= C(X, Y) + 0.25S(X, Y) + R(X, Y), \\ \mathcal{D}_f(X, Y) &= C(X, Y) + 0.25S(X, Y) + 2R(X, Y) + N(X, Y). \end{aligned} \tag{8}$$

The event-statistic term E averages reconstruction Smooth-L1 losses over sites, and $\bar{g}$ averages their gate activations.

For clean RGB with its associated clean-condition events, let H_{id} denote the repaired final output and $\bar{g}_{\mathrm{id}}$ its mean gate. The identity objective is

$$\mathcal{L}_{\mathrm{id}} = 2C(H_{\mathrm{id}}, H_c) + 0.5S(H_{\mathrm{id}}, H_c) + R(H_{\mathrm{id}}, H_c) + N(H_{\mathrm{id}}, H_c) + 0.1\bar{g}_{\mathrm{id}}. \tag{9}$$

This learned clean-input identity constraint differs from the exact empty-event identity in Equation 4.

For degraded RGB, let H_+ use matched events and H_- use mismatched events. The ranking objective is

$$\mathcal{L}_{\mathrm{rank}} = 2\left[C(H_+, H_c) - C(H_-, H_c) + 0.1\right]_+, \tag{10}$$

where $[a]_+ = \max(a, 0)$. Negative events come from the previous microbatch on the same rank when shapes match. Otherwise, we circularly shift the current voxel tensor by one-third of its height and width. Both repaired branches receive gradients. A microstep processes one example per GPU before gradient accumulation; every second microstep includes identity and ranking as specified in Equation 6, without additional frequency compensation. The active objective has no pixel loss, token-contrastive loss, or language-hidden-state loss.

B.4 Inference and the supervised driving study

SDE inference uses the Qwen chat template, one image and one question per request, greedy decoding, and at most 128 generated tokens. The prompt requests one JSON object with the sole key `answer`, without explanations or Markdown. The answer type is Boolean, non-negative integer, concise string, full multiple-choice text, or ordered text-string list. All Qwen conditions and placement controls use `sde_complete_literal_v3` normalization followed by typed validation; remaining invalid answers receive zero. The decoder remains shared within the Qwen3 comparison. Appendix H specifies the separate Qwen2.5 reproduction and scoring update; Appendix F defines InternVL's normalization.

The EventDrive comparison in Table 7 trains Qwen2.5-VL-7B adapters from scratch on the six official training files, totaling 249,539 rows. Both settings use answer-token cross entropy, sites $[7, 15, 23, 31]$, 560×420 RGB inputs, and 300 visual tokens. The objective excludes clean-representation alignment, parameter-anchor regularization, and mismatched-event losses. Training uses neither synthetic degradation nor event dropout. Both settings freeze the original vision and language weights and the auxiliary event decoders; joint adaptation additionally trains language LoRA.

Each setting uses two GPUs, batch two per GPU, accumulation 32 (effective batch 128), seed 42, and two epochs totaling 3,900 optimizer updates. AdamW uses adapter learning rate 10^{-4}, zero weight decay, 100 warmup updates, cosine decay, and gradient clipping at one. We use bfloat16, gradient checkpointing, and maximum sequence length 2,048. The joint setting uses LoRA rank 32, scaling parameter 64, dropout zero, and learning rate 10^{-5} on the language model's query, key, value, output, gate, up, and down projections. Final checkpoints serve both Regular and Hard; neither test split supplies training labels or checkpoint selection.

Driving inference uses the released task scorers after choice-format normalization. For adapter-only Regular trajectory scoring, deterministic normalization rewrites four explicit `0+number` coordinates as numeric literals, retaining their values and requiring exactly ten coordinate pairs. All 9,842 trajectories then parse; the resulting mean L2 is 7.2022 meters. The archived audit retains the four original strings and their normalized forms. Appendix E compares the main adapter-only setting with the early-site configuration; both driving settings remain separate from SEE-to-SDE representation training.

B.5 Public-code Event-MLLM comparison

We retrain the public Event-MLLM alignment-stage architecture on SEE-600K with Qwen3-VL-2B (Zhang et al., 2026). Frozen Qwen vision passes encode Degraded RGB and a rendered event image; a frozen DINOv2-B branch (Oquab et al., 2024) supplies illumination features. This baseline trains an 18,358,272-parameter merger (18.36M) to combine these signals and match clean final tokens with mean squared error. ECLAR separately trains 4,892,972 parameters (Appendix B.2). Fusion replaces only final visual tokens; DeepStack retains the Degraded RGB intermediate tokens. The experiment excludes instruction tuning and language model LoRA.

The implementation follows public commit `2d7d3cd0a168`. It renders five-bin event voxels as red/blue polarity images and resizes inputs to 336×336 at training and inference. We train for three SEE epochs with Adam, constant learning rate 10^{-3}, effective batch size four, and seed 3407. Two GPUs use accumulation two. A 500-pair SEE validation set selects the checkpoint by representation loss. Table 2 labels this result as public-code retraining on SEE; it evaluates the adapted alignment stage, with its own preprocessing and optimization recipe.

B.6 DA-CLIP RESTORATION CASCADES

We evaluate DA-CLIP (Luo et al., 2024) using the official ViT-B/32 controller and universal image restoration weights from code commit `c2cc14146faa`. The official variant uses the released weights directly. The +SEE variant freezes the controller and fine-tunes the IR-SDE Conditional U-Net on the same 171,711 SEE training pairs as ECLAR. Training retains the released matching objective, with AdamW, initial learning rate 2×10^{-5}, cosine decay to 10^{-6}, batch size 16, 256×256 patches, flip/rotation augmentation, and seed 3407. The 32,196-step budget corresponds to approximately three sample passes with replacement. We evaluate the final checkpoint; eight held-out SEE pairs monitor training, without SDE-based checkpoint selection.

Both variants restore all 1,838 SDE frames at their original resolution using 100-step posterior sampling, maximum noise level 50, a cosine schedule, and $\epsilon = 0.005$. Per-image seeds match across variants. The controller uses its native 224-pixel preprocessing; the restoration output retains the full image. Inference uses degraded RGB alone, with no events or clean targets. The shared Qwen loader applies orientation handling once and passes restored RGB to the frozen VLM without activating repair adapters. Both variants cover the same 8,357 question identifiers with no duplicates or missing predictions and use the common `sde_complete_literal_v3` answer parser. These are VLM-guided restoration cascades, not restorers trained with downstream Qwen-token losses.

C METRIC DEFINITIONS AND PAIRED UNCERTAINTY

Attribute, Counting, Object Recognition, and Scene Understanding use accuracy under the deterministic answer contract. Object Presence and Text Presence use balanced accuracy, the average of the two class recalls. For OCR Transcription, the scorer normalizes strings and averages $1 - d_{\text{edit}}(a, b)/\max(|a|, |b|, 1)$ over aligned entries in the predicted and reference text lists. A list-length mismatch receives zero. This ANLS implementation uses no additional similarity threshold.

Macro first combines Text Presence and OCR Transcription into one equally weighted OCR family, then averages the six family scores. Let A, C, P, R, S, T, O denote the seven task scores in Table 1(b), in that order. Then

$$\text{Macro} = \frac{A + C + P + R + S + (T + O)/2}{6}. \tag{11}$$

For example, InternVL clean RGB has OCR family $(50.61 + 12.69)/2 = 31.65$ and Macro $(78.88 + 45.35 + 79.84 + 80.96 + 81.26 + 31.65)/6 = 66.32$ after rounding. Micro is $\frac{1}{8357} \sum_{q=1}^{8357} s_q$, where s_q is the raw per-question score. Boolean questions use ordinary correctness before class balancing; OCR questions use normalized edit similarity. Thus Micro weights individual questions, not the seven displayed task scores equally. All tables multiply scores by 100; contrasts use unrounded values.

The Macro recovery fraction is $(M_{\text{repair}} - M_{\text{low}})/(M_{\text{clean}} - M_{\text{low}})$, where M denotes Macro. Qwen3 gives $(0.668558 - 0.549457)/(0.757579 - 0.549457) = 57.2265\%$. This fraction uses Macro alone.

We quantify the main configuration's Macro gain over Degraded RGB with a paired cluster bootstrap. Each draw samples 91 sequence identifiers with replacement and retains all questions within each sampled sequence, including multiplicities. The scorer recomputes task accuracies, class-balanced accuracies, OCR scores, and the six-family Macro for both conditions. The empirical 2.5th and 97.5th percentiles of 10,000 paired differences form the 95% interval. The reporting script uses base seed 3407.

For $[1, 2, 3, 4]$, the consolidated report yields an 11.9101-point gain and interval $[9.2937, 14.4204]$. The interval conditions on the evaluated checkpoint and layout; it excludes training-seed variation

and architecture-selection uncertainty. It is not an interval for the contrast against SEE-Net, the RGB-only control, or another injection layout. The script sorts condition names and adds each condition's index to base seed 3407; the archived report fixes that condition list.

D TASK-LEVEL RESULTS

Table 8 separates the seven question types underlying the main aggregates. Table 9 provides the family-level breakdown for all six layouts. These tables use the same reviewed question set as the main comparisons.

Table 8: Seven question-type scores for the Qwen3 paired SDE comparison. Presence columns use balanced accuracy; OCR uses normalized edit similarity.

Condition	Attr.	Count	Object presence	Recognition	Scene	Text presence	OCR
Clean RGB	84.16	44.63	88.92	81.95	85.25	83.05	56.24
Degraded RGB	72.50	13.60	54.66	74.69	64.75	58.27	40.68
EvLight	67.38	13.13	66.26	64.21	52.30	70.54	43.80
RetinexFormer	68.72	22.91	72.33	71.34	67.05	74.73	48.07
SEE-Net	72.58	28.64	75.65	71.65	64.39	75.31	46.53
DA-CLIP (official)	70.76	29.59	78.38	70.10	67.23	75.40	44.49
DA-CLIP (+SEE)	77.31	23.63	73.27	71.22	70.44	71.07	41.49
Event-MLLM (SEE retraining)	74.15	23.63	68.03	66.44	57.50	55.31	4.33
RGB-only control	75.26	33.41	79.09	70.97	71.16	73.29	38.25
ECLAR $[1, 2, 3, 4]$	75.97	40.57	82.41	74.38	71.46	74.03	38.64

Table 9: Semantic-family breakdown of the six injection layouts. The OCR family averages Text Presence and OCR Transcription.

Injection blocks	Attr.	Count	Presence	Recognition	Scene	OCR family
$[5, 11, 17, 23]$	76.20	33.89	81.92	72.58	71.52	55.61
$[4, 10, 16, 22]$	75.49	31.98	83.31	73.57	74.49	55.44
$[4, 10, 16, 23]$	76.04	35.32	82.10	72.95	71.04	56.11
$[1, 2, 3, 5]$	74.78	36.75	82.94	75.81	71.16	57.10
$[1, 2, 3, 4]$	75.97	40.57	82.41	74.38	71.46	56.34
$[0, 1, 2, 3]$	74.55	40.33	83.62	73.51	72.43	55.90

E DRIVING CONFIGURATION COMPARISON

Table 10 compares two adapter-only configurations. D denotes the main distributed sites $[7, 15, 23, 31]$; E denotes early sites $[1, 2, 3, 4]$. Both register 5,550,380 event-encoder and repair-block parameters and freeze the VLM. D freezes auxiliary decoders and trains 5.13M parameters; E trains all 5.55M (Appendix B.2). Both use Qwen2.5-VL-7B, 300 visual tokens, random adapter initialization without SEE pretraining, and two epochs over 249,539 EventDrive training rows.

D repairs states after global attention; E acts before the first global block (Bai et al., 2025b; Qwen Team, 2025b). Event tokens and statistics follow the RGB window-index permutation, supporting patch-aligned residual fusion in both layouts. The comparison therefore varies both cross-window context and depth coverage.

Both use effective batch 128 and 3,900 optimizer updates. D follows the pure-SFT recipe in Appendix B.4. E uses seed 42, two GPUs, batch two per GPU, accumulation 32, and adapter learning rate 10^{-4}. Its objective includes a parameter-anchor penalty with weight 10^{-3} and no mismatched-event training term. The different objectives and trainable auxiliary heads prevent attribution of the score differences to placement alone.

Table 10: EventDrive adapter-only configuration comparison. Objectives and trainable auxiliary heads differ. Scores are percentages except trajectory L2 in meters. Grounding uses the original-coordinate xywh contract.

	Regular		Hard	
Metric	D	E	D	E
Trainable parameters (M)	5.13	5.55	5.13	5.55
Perception: DSEC	92.56	71.25	94.36	63.08
Perception: M3ED	94.50	76.03	94.31	74.94
Understanding: QA	57.12	50.04	61.46	46.62
Grounding: Acc@0.6	5.82	0.54	4.64	0.00
Grounding: mIoU	22.50	9.50	23.90	8.40
Prediction: Speed	44.91	14.31	51.27	29.24
Prediction: Path	92.17	19.22	81.22	34.62
Planning: Speed	73.05	50.13	76.14	65.73
Planning: Path	87.39	86.77	90.50	83.17
Trajectory: L2↓	7.2022	8.29	6.5968	7.46

Metric coverage and interpretation. For the metrics in Table 10, the early regular reports score 6,414 DSEC and 16,068 M3ED perception questions, 12,020 understanding questions, 2,404 grounding boxes, 1,405 prediction questions, and 9,842 planning examples. Hard contains 390, 7,698, 755, 151, 985, and 6,317 respectively. The perception and understanding scorer combines the option-letter and label-text responses; prediction and planning report speed and path separately. These aggregate accuracies do not establish balanced class recall. We retain each report's mIoU precision: its three-decimal IoU becomes a percentage with two decimal places.

Coordinate conventions. Grounding requires original-image coordinates in (x, y, w, h) form, while preprocessing resizes RGB to 560×420. The early adapter-only outputs show sensitivity to coordinate format and scale. On its 2,404 regular outputs, a fixed diagnostic interpretation as input-canvas corner coordinates yields 42.35 Acc@0.6, compared with 0.54 under the recorded contract. Table 10 retains 0.54; the diagnostic conversion does not replace formal scoring or measure a change in model weights. Grounding differences therefore combine localization with output-contract adherence.

F INTERNVL2.5-2B PROTOCOL AND TASK DETAILS

Task details complement the main comparison. Table 11 separates Text Presence and OCR Transcription for InternVL2.5-2B (Chen et al., 2024). All conditions cover the same 8,357 identifiers. Counting and Object Presence gain 15.99 and 10.95 points over Degraded RGB.

Table 11: Task-level detail for InternVL2.5-2B, complementing Table 2(b). Object/Text presence use balanced accuracy; OCR uses normalized edit similarity.

Condition	Mod.	Attr.	Count	Obj. pres.	Recogn.	Scene	Text pres.	OCR	Macro	Micro
Clean RGB	R	78.88	45.35	79.84	80.96	81.26	50.61	12.69	66.32	65.99
Degraded RGB	R	**71.32**	19.81	62.42	69.23	68.38	53.62	13.88	54.15	56.94
Restoration cascades										
EvLight	R+E	62.33	25.78	63.43	59.99	49.76	48.63	12.23	48.62	48.99
RetinexFormer	R	63.12	32.22	68.02	67.87	63.24	**55.53**	**15.84**	55.03	56.94
SEE-Net	R+E	65.72	34.84	65.83	69.04	58.28	46.01	12.35	53.82	54.38
ECLAR [1, 2, 3, 4]	R+E	71.00	**35.80**	**73.37**	**72.83**	**69.17**	51.34	7.03	**58.56**	**59.23**

The adapter preserves the native visual interface. We train a new SEE-600K adapter with InternVL's vision encoder, projector, and language model frozen. Sites $[1, 2, 3, 4]$ update patch tokens while retaining the class token, pixel shuffle, and projector. Full-frame 448×448 resizing supplies a 32×32 patch grid and 256 language-side tokens, with one view throughout training and all six inference conditions. Width 1024 gives 4,892,972 trainable parameters. Bypass and empty-event smoke tests reproduce native visual outputs.

Training uses Appendix B's objective and coefficients: three SEE epochs, AdamW at 2×10^{-4}, weight decay 0.05, 500 warmup updates, and cosine decay. Four GPUs, one example per rank, and accumulation two give effective batch eight; seed is 3407. Evaluation uses final update 64,392 without SDE epoch selection or Qwen-trained adapter weights.

One answer protocol applies to all six conditions. Inference shares the native InternVL chat template, questions, answer schema, greedy decoding, and 128-token limit. Deterministic normalization extracts complete, unambiguous literals using answer types and choices, never correct answer values; typed validation assigns remaining invalid answers zero. The archived protocol is `internvl_literal_serialization_v2`; earlier exports are historical. Invalid-answer counts are 95, 59, 89, 108, 100, and 84 for clean RGB, EvLight, Degraded RGB, ECLAR, RetinexFormer, and SEE-Net. Preprocessing and normalization differ from Qwen; claims compare conditions within each backbone.

Macro gains are 4.4078 points over Degraded RGB (10,000-draw paired 91-sequence bootstrap 95% interval: $[1.9574, 6.9071]$) and 3.5340 over RetinexFormer (no reported interval). Degraded RGB and RetinexFormer both round to 56.94 Micro; their scores before table rounding are 56.935758% and 56.935324%, respectively. Table 2(b) ranks these values before rounding; this difference supports no substantive performance claim. This single-seed experiment fixes the early layout, without InternVL-specific placement search or capacity-matched RGB-only training.

G Fine-Grained Grounding: Resolution, Residuals, and Alignment

Fixed-checkpoint probes separate spatial effects. Grounding requires spatial precision beyond a categorical answer. We examine resolution, repair-residual strength, and input alignment using an earlier anchor-regularized adapter-only Qwen2.5-VL-7B checkpoint at $[7, 15, 23, 31]$. This diagnostic checkpoint differs from the pure-SFT checkpoints evaluated in Table 7. The probes characterize its DSEC grounding behavior under explicit diagnostic protocols; they establish no cause for the new checkpoints' grounding, trajectory, or OCR results.

G.1 Resolution helps under a unified coordinate request

Protocol separates geometry from output formatting. We fix 64 examples from a previously sampled regular-split grounding subset, preserving the referring expressions across conditions. The diagnostic prompt requests corner coordinates (x_1, y_1, x_2, y_2) on the actual model input canvas. The parser extracts the first four numbers and deterministically converts them to original-image (x, y, w, h) coordinates before IoU scoring. This measures extractable geometry rather than strict four-number compliance: native Qwen often emits additional numbers. The subset and protocol differ from Table 7; we use neither to choose a reported model.

Table 12: Resolution and repair on 64 fixed grounding examples under the diagnostic coordinate protocol. Scores are percentages; tokens count language-side visual tokens.

Model	Actual RGB input	Tokens	Acc@0.6	mIoU
Native Qwen	560×420	300	79.69	71.04
ECLAR	560×420	300	6.25	19.48
Native Qwen	644×476	391	89.06	76.45
ECLAR	644×476	391	37.50	42.79

Input size explains part of the diagnostic gap. Passing the original 640×480 image directly to the processor produces a 644×476 input and 391 visual tokens. Both models improve under the unified request (Table 12), with event input fixed at 346×260. However, native Qwen already retains localization at the smaller size, and the repaired model remains behind it at both sizes. The intervention changes the token grid and coordinate canvas as well as pixel detail. Under the original published xywh prompt on 128 examples, increasing input size changes repaired Acc@0.6 from 8.59 to 5.47, so resolution alone is not a protocol-independent remedy.

G.2 Full-strength residuals interfere with grounding

Closing the repair path restores native outputs. At each repair site, we replace the realized output by $z + \beta(\hat{z} - z)$ without changing learned weights. We use one additional inference multiplier β across all sites; the learned layer-specific α_ℓ remains fixed. All other inputs and decoding settings match the 64-example fixed-resolution probe. Both bypass and executing repair blocks with zero gates reproduce the native model's 64 answers exactly. On eight examples per condition, the visual outputs and next-token logits also match the unwrapped native model with maximum absolute difference zero. These checks isolate the active repair residual from loading or wrapping side effects in the inspected paths.

Table 13: Fixed-checkpoint inference interventions, without retraining: (a) residual strength on 64 examples; (b) single-site activation on the first 32.

(a) Residual strength ($n = 64$)

Condition	Acc@0.6	mIoU
Native / bypass	79.69	71.04
Zero gates	79.69	71.04
$\beta = 0.10$	75.00	68.98
$\beta = 0.25$	76.56	69.84
$\beta = 0.50$	62.50	57.53
$\beta = 1.00$ (full)	6.25	19.48

(b) Active sites ($n = 32$)

Condition	Acc@0.6	mIoU
None	84.38	71.62
Only 7	84.38	70.61
Only 15	81.25	70.39
Only 23	81.25	65.65
Only 31	71.88	65.37
All four	9.38	22.55

Interference depends on residual strength and composition. Smaller residuals recover much of the extractable localization, although the strength sweep is not strictly monotonic (Table 13). Each site alone retains more grounding accuracy than all four together. These inference interventions disable sites in one jointly trained driving checkpoint; they do not compare separately trained single-site and multi-site models on SDE. The full model's mean gates at sites 7, 15, 23, and 31 are 0.948, 1.000, 0.999, and 0.999; residual-to-input root-mean-square ratios range from 5.71% to 6.55%. Gate saturation describes the operating point; the interventions provide the evidence of residual interference. They leave the upstream training cause unresolved and do not establish a residual scale that improves every task.

Box-center errors dominate this diagnostic protocol. Full repair increases median center distance from 4.95 to 47.59 pixels in original-image coordinates. Median horizontal/vertical errors are $+34.21/+20.14$ pixels; width/height ratios remain 1.052/0.951. Replacing predicted centers with ground truth yields 63.57 mIoU, versus 20.39 for replacing sizes. These are diagnostic interventions, not deployable corrections. A two-fold translation correction, fitted on one half and evaluated on the other, reaches 34.88 mIoU: the error exceeds a constant offset. Original xywh prompts yield a different profile, so predicted-box shifts do not establish an RGB–event calibration offset.

G.3 Input alignment requires branch-specific checks

Real capture does not ensure pixel correspondence. We inspect the DSEC and M3ED branches of EventDrive (Lu et al., 2026). The DSEC release remaps RGB into the event view and distinguishes event windows before or after an image (Gehrig et al., 2021; Gehrig & Scaramuzza, 2024); raw M3ED provides calibration and temporal indexing (Chaney et al., 2023). The audit covers 56

released pairs: 32 DSEC grounding images from seven sequences and 24 M3ED samples from three task groups. All paired filenames agree; this verifies neither calibration nor exposure timing.

DSEC and M3ED share release canvases. The inspected DSEC pairs use 640×480 RGB and event arrays; M3ED pairs use 640×400. In DSEC, an event-activity-to-RGB-edge translation search gives a componentwise median displacement of $(-1, 0)$ pixels. The across-image median capped edge distance changes from 2.14 to 2.03 pixels after the search. This texture-dependent proxy, together with visual inspection, gives no evidence of a uniform input displacement of tens of pixels. Sparse events make several estimates unstable, particularly for M3ED. These proxies are not calibration ground truth. The inspected dense-event files lack absolute event-window and exposure metadata, preventing a complete temporal-alignment check.

Training hypotheses remain open. Grounding supplies 2,739/249,539 training rows (1.10%). We trained the diagnostic checkpoint with answer-token cross-entropy and a parameter anchor, without free-generation grounding validation or explicit spatial preservation. These diagnostics establish no unique cause or remedy. Testing task balance and coordinate supervision requires controlled retraining with independent grounding and categorical-task validation.

H Official-Data Event-MLLM Reproduction

Original-task validation supports comparable reproduction. We reproduce Event-MLLM (Zhang et al., 2026) with Qwen2.5-VL-3B, the released data, and both training stages. The released validation set contains 160 multiple-choice and 114 counting questions, each evaluated at 17 brightness ratios. We retain the official inference and scoring routines, 336×336 inputs, greedy decoding, seed 42, bfloat16, and the 512-token generation limit. Multiple-choice accuracy measures exact option-set matches; F1 averages the 17 brightness-specific micro-F1 values. Counting accuracy and mean absolute error (MAE) average the six object classes. These metrics differ from SDE-VQA-Bench's single-answer task scores.

Table 14: Event-MLLM original-task validation with Qwen2.5-VL-3B. Published values follow Zhang et al. (2026), Table 2. Accuracy and F1 are percentages.

Metric	Published full model	Our reproduction
Multiple-choice accuracy	34.48	32.90
Multiple-choice F1	80.35	79.78
Counting class accuracy	71.08	72.06
Counting class MAE ↓	0.6127	0.5834

Table 14 places the reproduction near the published performance: multiple-choice accuracy differs by -1.58 points and F1 by -0.57, while counting accuracy improves by 0.98 points and MAE decreases by 0.0293. This validates comparable original-task performance without implying numerical equivalence.

Released-code settings define the training recipe. Stage 1 trains for 30 epochs with batch size two and learning rate 10^{-4}, following the pinned released configuration. The paper instead describes batch size four and learning rate 10^{-3}. Stage 2 applies one epoch of language LoRA instruction tuning. Following the released renderer's first-frame handling, preprocessing excludes 12 DSEC pairs, corresponding to 72 questions; 10,056 of 10,128 training questions remain. Validation retains all released questions. Inference strictly loads the best Stage 1 merger and the complete final Stage 2 state. This experiment uses locally trained reproduction weights, distinct from the SEE-retrained alignment stage in Appendix B.5.

SDE evaluation uses real low-light observations. Table 2(c) evaluates all 8,357 reviewed questions using Degraded RGB and associated raw events, with no synthetic brightness scaling or clean-image input. We retain the official event renderer, correct color order to match its image I/O path, and apply the benchmark's orientation metadata to both streams. Inputs use 336×336 resizing and

generation uses at most 128 tokens. Text Presence scores 70.04 and OCR Transcription 10.54, giving OCR-family 40.29. The training corpus includes SDE sources, so this evaluation does not establish held-out cross-dataset generalization. Source overlap alone does not quantify exact benchmark-frame overlap. Backbone and training differences also preclude attributing its contrast with Qwen3 systems solely to fusion design.

Shared normalization separates answer content from serialization. All Qwen conditions, including the RGB-only control, six layouts, and encoder substitutions, use one answer-blind normalizer followed by unchanged typed validation and scoring. Normalization uses only the requested answer type and supplied choices, never the correct answer value; it rejects ambiguous or conflicting outputs. Raw generations remain unchanged. The official-data model supplies 8,353 valid answers among 8,357 questions. InternVL retains its shared within-backbone protocol from Appendix F.

I EVENT-ENCODER SUBSTITUTION

A shared recipe tests encoder compatibility. ResNet-18 (He et al., 2016), EFNet-derived (Sun et al., 2022), RVT-T (Gehrig & Scaramuzza, 2023), and Spikformer (Zhou et al., 2023) share the five-bin, two-polarity input, 256-dimensional event interface, 320-dimensional bottleneck, and sites $[1, 2, 3, 4]$. All start randomly with Qwen3-VL-2B frozen and follow Appendix B's SEE data, losses, and three-epoch recipe: one GPU, batch one, accumulation eight, seed 3407. Final checkpoints answer all 8,357 SDE questions without SDE checkpoint selection. The main encoder retains its existing run; RVT-T and Spikformer resume at update 5,000 with optimizer and scheduler states restored.

Table 15: Seven-task SDE scores for event-encoder substitution, complementing Table 5. Presence columns use balanced accuracy; OCR uses ANLS.

Event encoder	Attr.	Count	Obj. pres.	Recogn.	Scene	Text pres.	OCR
Lightweight CNN	75.97	40.57	82.41	74.38	71.46	74.03	38.64
ResNet-18	74.94	33.65	78.79	71.15	70.07	72.47	35.10
EFNet-derived	71.16	28.88	77.03	69.11	64.21	73.05	36.67
RVT-T (adapted)	74.00	30.55	77.53	71.03	69.04	73.17	26.82
Spikformer (adapted)	74.86	32.46	80.65	70.97	69.35	73.34	36.40

Substitutions retain the repair interface. ResNet-18 uses ten input channels and retains batch normalization; a $512 \to 256$ projection with GroupNorm replaces its pooling/classification head. The EFNet-derived branch uses the released event stem and three event blocks. We resize their 64/128/256-channel outputs to the deepest grid, concatenate them, and project $448 \to 256$ channels with GroupNorm. We exclude EFNet's RGB branch, cross-modal attention, and image decoder. Both substitutions preserve voxel-statistics targets, patch alignment, and subsequent adapter initialization through an isolated encoder random-number stream.

Temporal encoders use the same event window. RVT-T retains four recurrent stages of widths $[32, 64, 128, 256]$, omitting the detection neck/head. It processes five chronological polarity pairs and resets state per sample. Attention uses 4×4 partitions; right/bottom padding to multiples of 128 supports the hierarchy. The stride-32 output discards padding before projection to 256 channels. Spikformer retains the released dynamic vision sensor (DVS) patch stem and two 256-channel, 16-head spiking-attention blocks. Dynamic reshaping replaces fixed-square reshaping; temporal averaging preserves spatial output. Leaky integrate-and-fire neurons use PyTorch FP32 and reset per sample. Both retain the shared projection, GroupNorm, voxel-statistics targets, and patch alignment, without extra frames or cross-sample history.

Trainable totals, including auxiliary heads, are 15,951,404, 6,382,572, 7,879,212, and 7,251,500, respectively; the main model has 4,892,972. These single-seed substitutions share a nominal recipe without parameter matching or architecture-specific tuning. Stride, normalization, precision, and GPU/resume schedules differ; the results evaluate adapted encoders within ECLAR. Under Appendix H's normalization, 8,259, 8,283, 8,224, and 8,275 answers pass validation, respectively; each

denominator remains 8,357. All four pass real-sample gradient, zero-event identity, and token-shape checks.

J VISUAL TOKEN REPRESENTATION RECOVERY

Final visual tokens recover clean-image representations. ECLAR achieves the highest cosine similarity and linear centered kernel alignment (CKA) (Kornblith et al., 2019), and the lowest relative L2 and relation error, among the evaluated methods (Table 3). DA-CLIP (+SEE) ranks second on all four metrics. The comparison measures how closely each system's final visual tokens approach the corresponding clean RGB tokens, complementing the answer-level results in Table 2(a).

The protocol reuses the paired benchmark frames. We evaluate the 1,838 distinct frames from the same 91 sequences underlying the 8,357-question benchmark, assigning each frame equal weight regardless of its question count. These SDE frames remain separate from SEE training; Section 5 states the SDE-based layout choice. We retain the main $[1, 2, 3, 4]$ final checkpoint and the restored images from the corresponding VQA evaluations. Each method uses the same frozen Qwen3-VL-2B-Instruct backbone and native image processor. Checks enforce matching image orientation, spatial dimensions, visual grids, and token shapes across each clean/degraded pair. The evaluation compares final post-merger tokens entering the language model and excludes intermediate DeepStack branches from the reported averages. Visual forward passes use bfloat16; metric calculations use float32.

Four metrics measure complementary aspects of recovery. For each frame, let $X, Y \in \mathbb{R}^{N\times d}$ denote the evaluated and clean-reference token matrices. Cosine similarity averages the cosine between corresponding token vectors. Relative L2 averages $\|X_i - Y_i\|_2 / \max(\|Y_i\|_2, 10^{-6})$ over token positions. Linear CKA centers each matrix across positions, forms their linear Gram matrices, and divides their Frobenius inner product by the product of their Frobenius norms (Kornblith et al., 2019). Relation error takes the mean absolute difference between the two token-to-token cosine-similarity matrices, using up to 256 evenly spaced positions. We compute each metric per frame, then average over all frames. Restoration cascades and ECLAR retain their respective inputs and training objectives, as in the answer-level system comparison.