NEURIPS 2026 · POSTER SESSION

Learning to Synergize Textual and Visual Prompts for Fine-Grained Traffic Element Detection

Xiaoyang Bi Haowen Guo Caoshengzhe Xue Siyuan Li
Tao Xue Dong Liu Mengshi Qi

Tencent Map · Key Laboratory of Artificial Intelligence, Ministry of Education

THE BENCHMARK

MFT-150K

150,495
images

337,147
boxes

55
fine-grained classes

Dataset, code and the full benchmark

huggingface.co/datasets
/MFTED/MFTED

Closed-set detection breaks when the map keeps changing

HD-map maintenance encodes traffic rules at lane level. A missed sign is not a missed box — it is a wrong map.

Every map update introduces regional, temporary or rare elements that were never in the training label set.

A closed-set detector needs new annotations and retraining for each new label, so cost grows with the long tail it is supposed to cover.

And the distinctions are **semantic**, not visual: an arrow painted on the road versus an arrow printed on a sign; a red slash that inverts the whole meaning.

CLOSED SET

A fixed label list. New class → new annotations → retrain.

QUERY-BASED

Free-form text plus visual exemplars. Detect what you can describe.

Question: can we detect by describing and pointing, instead of by naming?

MFTED: detect by describing and pointing

Given an image I and a query Q = (T , V), return every box that matches Q.

TEXT PROMPTS — FOUR KINDS OF SEMANTIC COMPLEXITY

Attribute visual characteristics that imply function — colour, shape, boundary

Relationship interplay with context — arrow on the road vs. arrow on a sign

Negation exclusion or prohibition — the red slash that flips the meaning

Reasoning infer the function from the symbol — which way does this sign lead

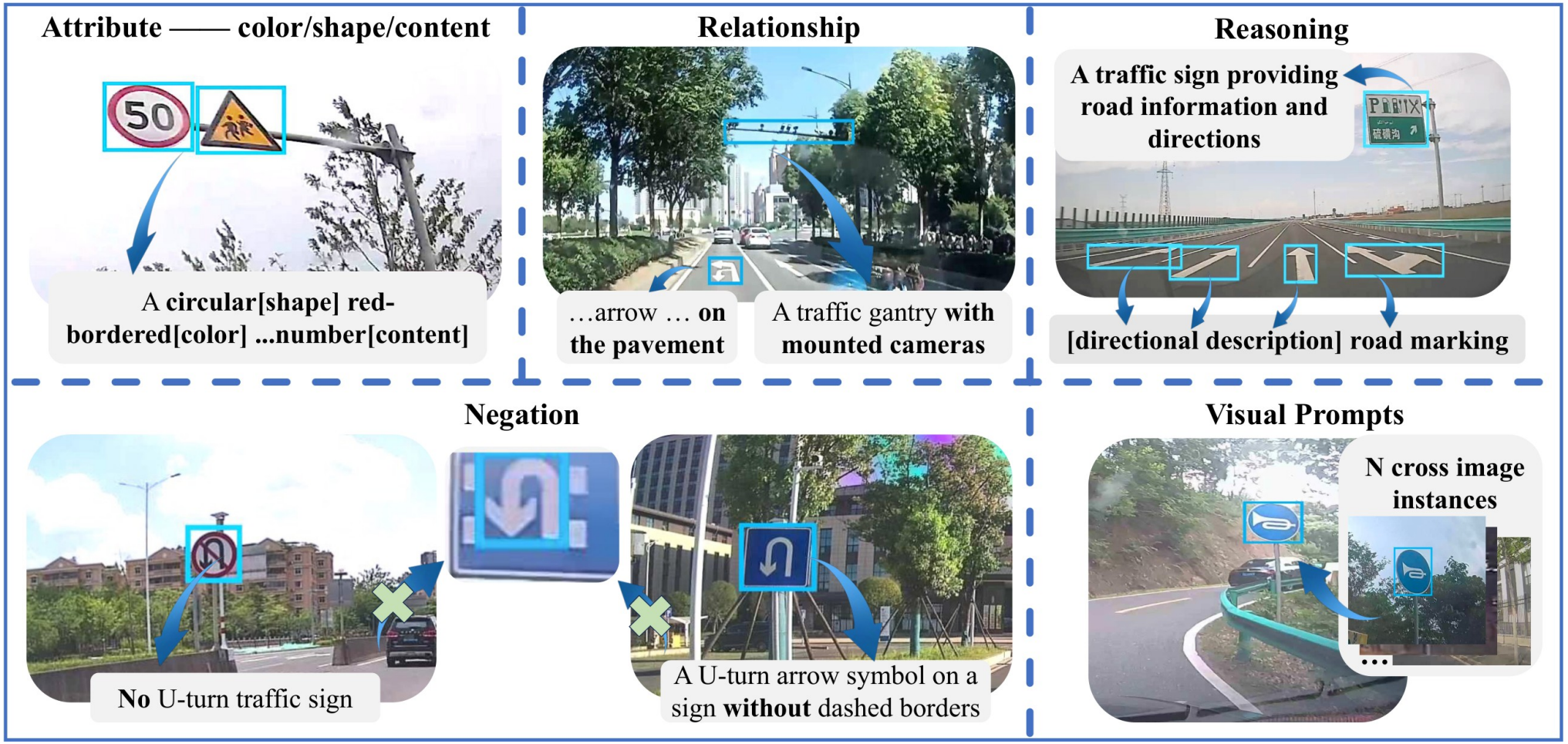
Query can be text only (T), visual only (V), or joint (G).

VISUAL PROMPTS

N reference images, each with a box annotation (default **N = 16**).

- Crop scale, lighting and viewpoint all vary.
- Exemplars are geographically disjoint from val/test — no scene leaks.
- They supply the grounding that text alone cannot.

A query is a text description plus visual exemplars



MFT-150K: 150K images, 337K boxes, 55 classes

150,495
images

337,147
box annotations

34 + 21
base + novel classes

31 / 300+
provinces / cities

Subset	Images	Test	Boxes	Test
Attribute	21,295	7,330	27,494	9,299
Relationship	50,836	12,993	66,716	17,064
Negation	8,639	3,302	9,162	3,478
Reasoning	82,647	22,719	173,774	45,611
MID	67,914	17,967	190,209	49,112
Novel	–	2,907	–	3,438
Total	**150,495**		**337,147**	

COMPOSITION

- Resolution: 1280×720 (48.6%), 1920×1080 (51.0%)
- Average description: **9.1 words** — not a class name
- Road mix: urban 57.3%, national / provincial 20.9%, expressway 11.0%, rural 10.9%
- **96.3%** of instances sit in a single semantic subset
- MID stress test: 67,914 images, 190,209 boxes

Annotated by mapping experts, verified by blind agreement

01

Expert authoring

HD-mapping experts write every description. No LLM generation.

02

Cross-validation

A second, independent expert reviews each prompt for ambiguity.

03

Spot-checking

Senior experts audit random samples against the boxes.

04

$\kappa = 0.92$

Three blind raters; almost perfect agreement.

Integrity by construction

Exemplars live in geographically disjoint regions from validation and test — no identical scene or sign is ever seen twice.

Faces, plates, timestamps and coordinates are redacted before release.

Built for deployment, not for a paper

The taxonomy and prompt style are already used in open-world perception for commercial autonomous fleets.

Released in full — all 150K images — under CC BY-NC-SA 4.0.

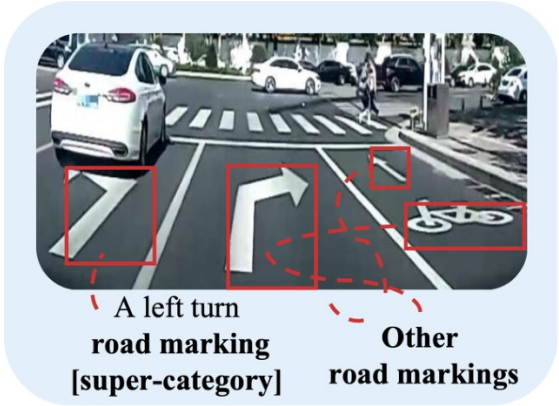
Base vs novel, plus a MID stress test

BASE → NOVEL SPLIT

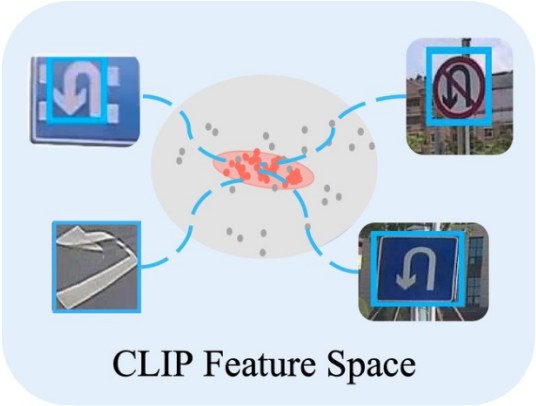
Train on 34 base classes, test on 21 disjoint novel classes ($C_{train} \cap C_{test} = \emptyset$). Zero-shot, training-free few-shot and fine-tuning are compared on the same N samples.

MULTI-INSTANCE DISCRIMINATION (MID)

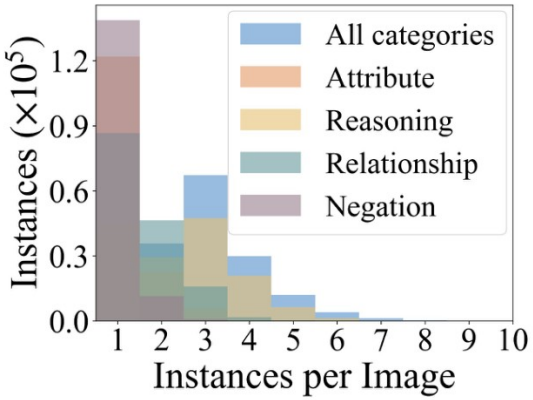
Co-located instances share a super-category but differ in label. Their CLIP embeddings cluster tightly, so keyword matching cannot separate them.



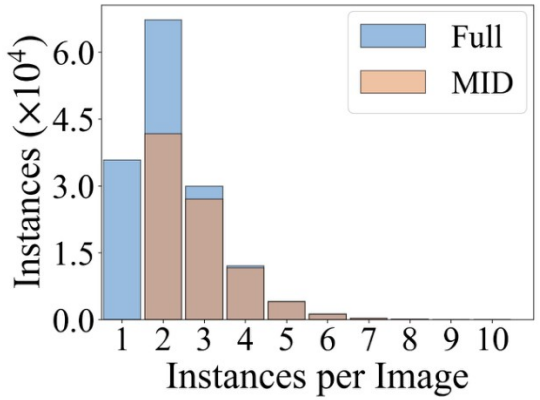
(a) MID



(b) Confusable Categories



(c) Instance Distribution



(d) Full vs. MID Distribution

Metric: mAP@[.5 : .95]. Regimes: **T** text only · **V** visual only · **G** joint prompts · with or without fine-tuning.

Naive fusion averages; we gate by modality reliability

WHAT EXISTING DETECTORS DO

$$P_f = f(P_t, P_v)$$

- T only, V only, or G with deterministic fusion (e.g. averaging).
- Equal reliability is assumed for every prompt.
- In fine-grained scenes that injects semantic noise instead of synergy.

Diagnosis: the fusion rule ignores prompt quality.

Text prompt T → P_t Visual exemplars V → P_v

↓

JVAE · factorised VAE into shared + private latents; σ per modality is read as confidence

↓

QAG · sigmoid gate on (σ_t , σ_v) up-weights the more reliable modality, element-wise

↓

P_f → detection head

Trained on base classes only; nothing in the gate is novel-class specific.

Off-the-shelf models collapse to under 10 mAP

Model	Pre-training data	Backbone	Params	Base	Novel
YOLOE	O365, GoldG	YOLOv8-L	26M	4.3	3.0
Grounding-DINO	O365, ALL	Swin-L	430M	0.8	0.6
SAM3	SA-Co	PE-L+	850M	9.2	7.1
Qwen-2.5VL-3B	–	–	3B	0.5	2.5

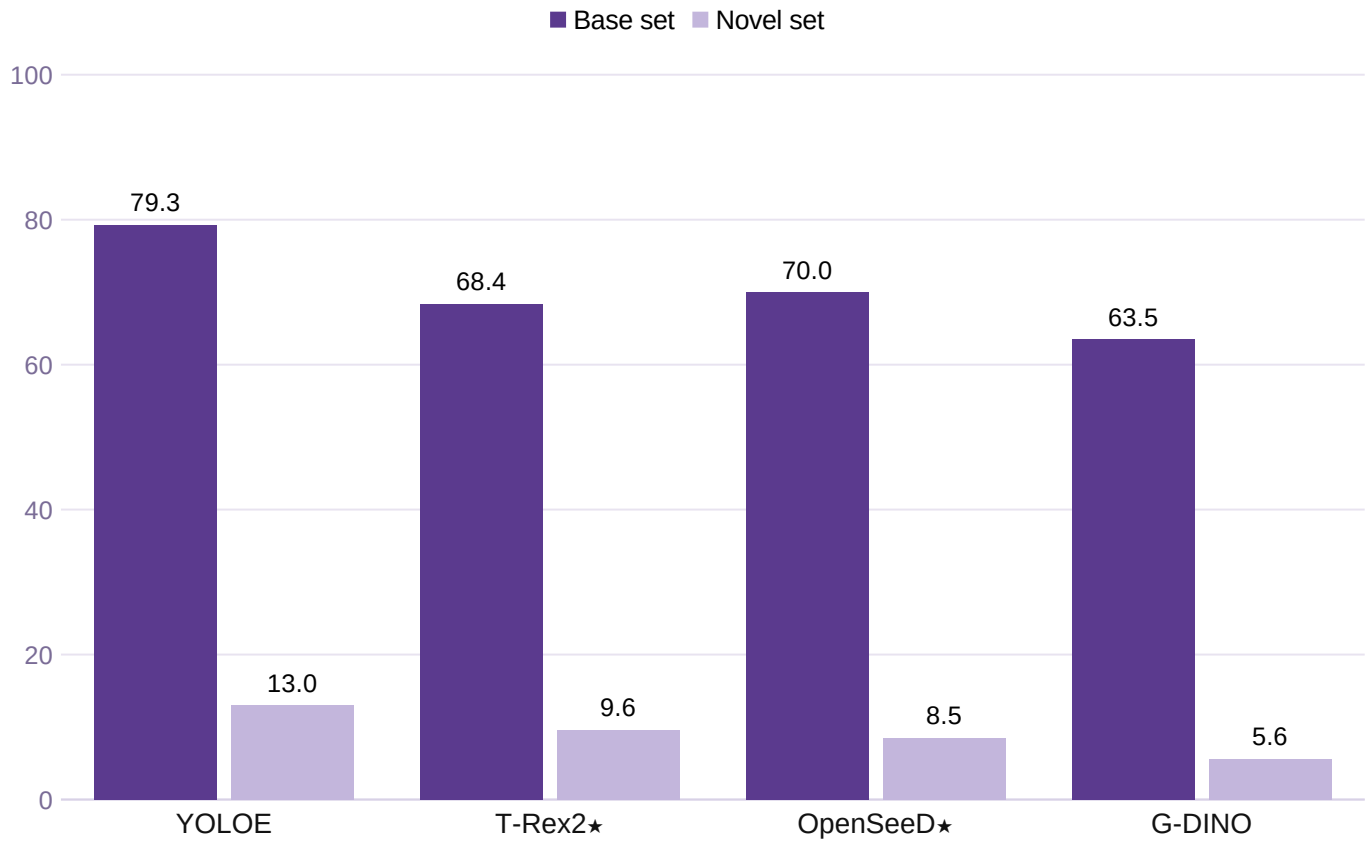
9.2 mAP is the best any model reaches — SAM3 with 850M parameters. Grounding-DINO, the standard open-vocabulary baseline, scores 0.8.

Open-vocabulary grounding does not transfer to traffic-rule semantics.

No fine-tuning on MFT-150K

Every model keeps its own large-scale pre-training, so this is the most generous zero-shot setting we can give them.

Text alone collapses; visual exemplars recover the novel set



Base → novel, zero-shot text

YOLOE drops from **79.3** to **13.0**. No amount of transcription accuracy survives a category that was never named.

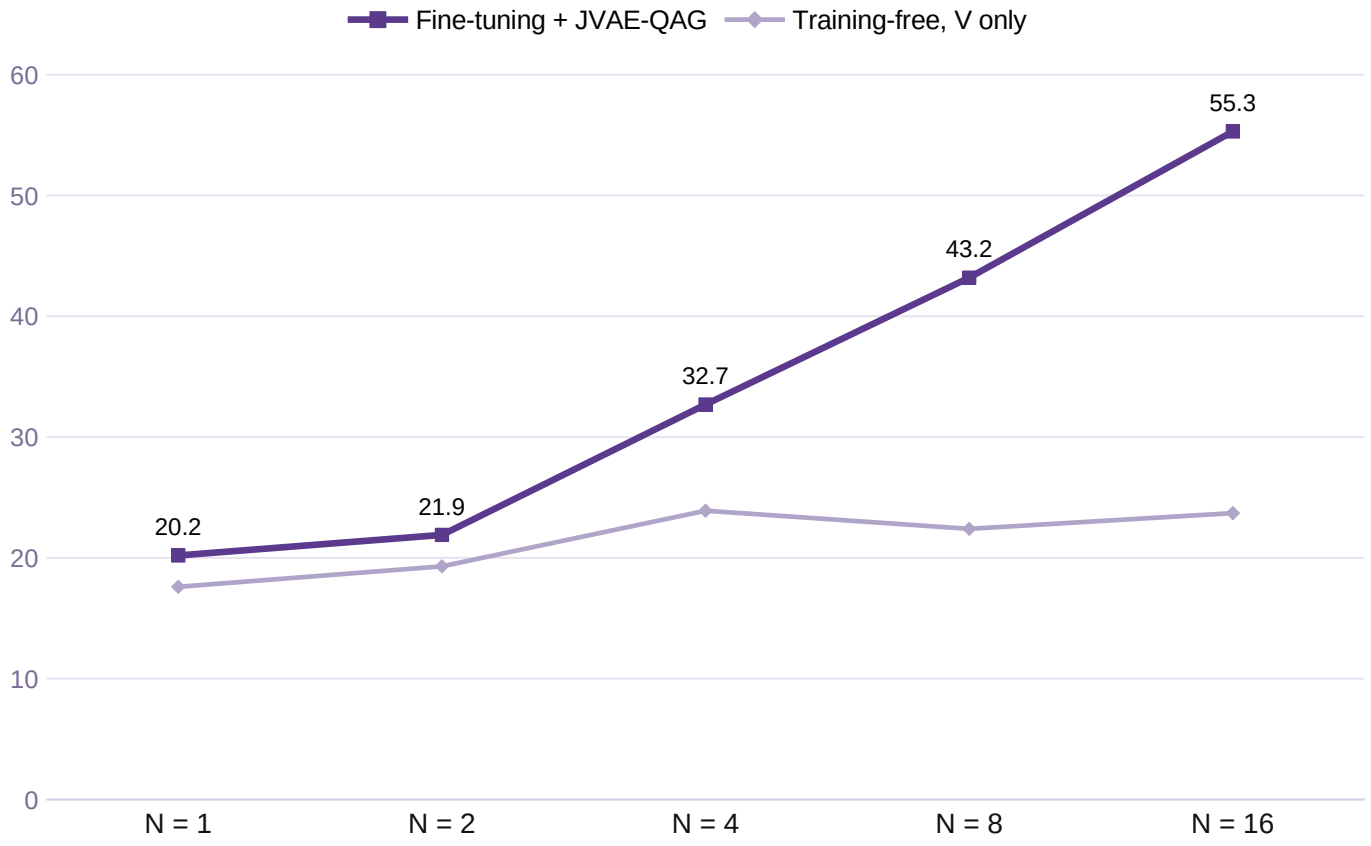
Visual exemplars, no gradients

Novel mAP rises **+10.1** for YOLOE and **+15.3** for T-Rex2 (Swin-T) — just from showing sixteen crops.

Joint prompts add nothing

T-Rex2 G scores 23.8 versus 23.7 for V alone: deterministic fusion is not fusion.

Fine-tuning, not fusion, moves the needle



Fine-tuning on 16 samples per class

Lifts FULL mAP from **19.7** to **55.3** — **+35.6** in total. This is the single largest effect in the paper.

What the two modules add

Fine-tuning +31.3, quality-aware gate **+3.5**, probabilistic alignment **+0.8**.

Training-free is a flat line

Twenty exemplars buy at most ~24 mAP. Without parameter updates, more examples do not help — the ceiling is the paradigm, not the data.

What actually helps: words, gating, the right category

16-shot ablation	MID	FULL	Δ
YOLOE	17.5	19.7	–
+ Fine-tuning	49.5	51.0	+31.3
+ QAG	52.7	54.5	+3.5
+ JVAE	**53.3**	**55.3**	**+0.8**

Prompt form	MID	FULL
Category name (C)	49.5	53.2
C + visual	49.9	54.0
Description (D)	52.4	54.0
D + visual	**53.3**	**55.3**

Words beat names

Replacing a class name with a description adds **+2.9** on MID; with visual prompts, **+3.4** and 55.3 FULL.

Fusion is category-conditional

+5.2 to +7.1 on directional and prohibitory signs, where arrow geometry is the signal; neutral or negative where language already suffices.

Gains by size: **+1.97** medium, **+2.67** large, **−0.36** small.

The gap is semantic, not lexical

Changing one element of a prompt while holding the image fixed costs **68–83%** mAP; negation is the sharpest.

A benchmark, a baseline, and an open problem

1 **The benchmark**

150,495 images and 337,147 boxes in 55 fine-grained classes, with base and novel sets that never overlap and a MID stress test built from confusable neighbours.

2 **The measurement**

Off-the-shelf VLMs stay under 10 mAP; text-only collapses to 13.0 on novel classes; fine-tuning on 16 samples reaches 55.3.

3 **The mechanism**

Quality-aware gating buys 3.5, probabilistic alignment only 0.8. Fusion is not the bottleneck — the semantics are.

OPEN PROBLEM

Training-free fusion on novel categories is still unsolved.

Frozen models plateau near 24 mAP however many exemplars you show them.

Dataset, code and full annotations

CC BY-NC-SA 4.0

huggingface.co/datasets/MFTED

A realistic benchmark, and a call for new fusion mechanisms.