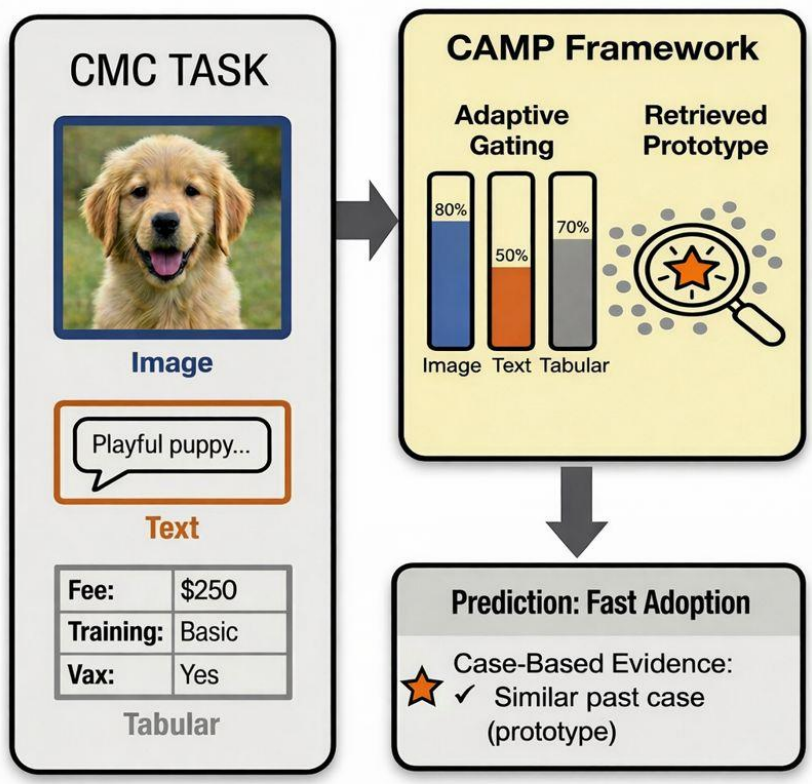
CAMP

Coherent Alignment of Multimodal Prototypes for Explainable Complementary Learning

A compact prototype head for complementary multimodal classification:
gate → retrieve → explain.

Alvaro Lopez Pellicer · Eoin M. Kenny · Simran Lamba · Shubham Sharma · Plamen P. Angelov · Saumitra Mishra

JPMorgan AI Research · Lancaster University



Lancaster University

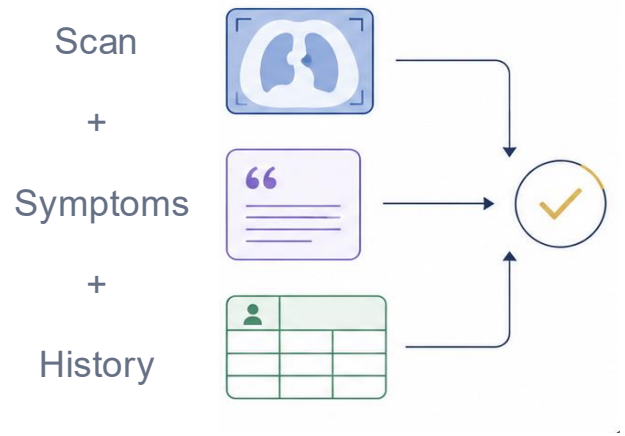
J.P.Morgan

POTENTIAL

CMC is perhaps the most practical real-world multimodal setting

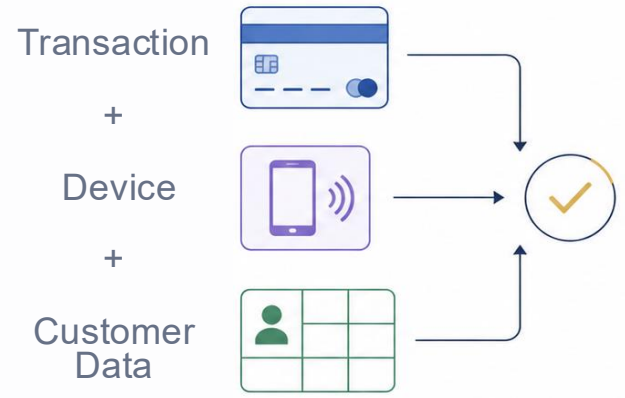
Important decisions often depend on evidence streams that complete one another rather than re-describe one another.

Clinical triage



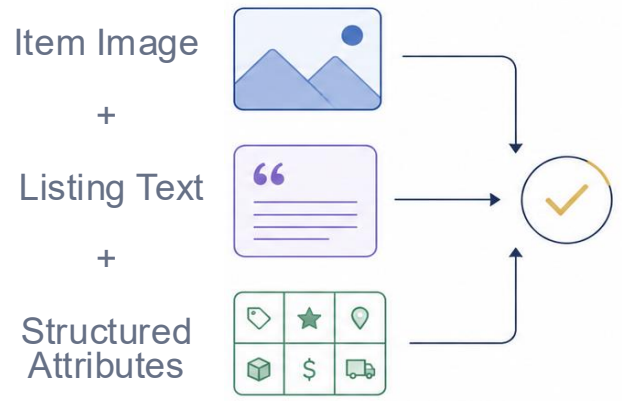
Why This diagnosis?

Fraud / risk



Which signal tipped the decision?

Marketplace decisions



Which similar market profiles fit this?

Needs: performance · auditable evidence · missing-stream robustness · case retrieval

Takeaway Real CMC cases are heterogeneous evidence bundles, not paired-view captioning tasks.

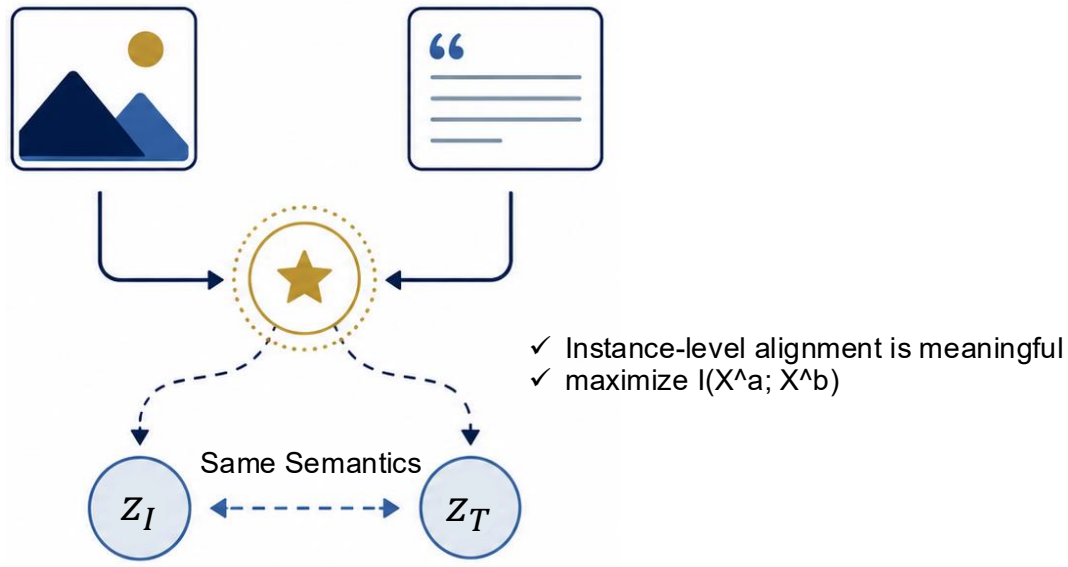
PROBLEM

Complementarity breaks paired-view alignment

In CMC, modalities can be weakly coupled but jointly predictive; forcing same-case embeddings together can erase the signal.

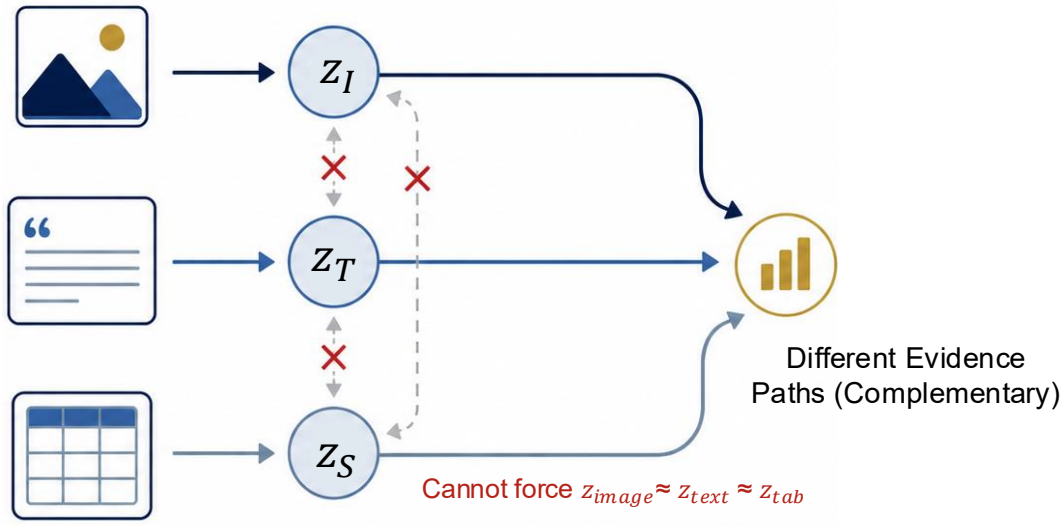
I(Y; X^(m) | X^(M \ {m})) > 0 — At least one modality adds label information given the others

Descriptive paired views



- ✓ Instance-level alignment is meaningful
- ✓ maximize I(X^a; X^b)

Complementary evidence



Note: z_m Refers to the embedding representation of a modality *m* which in this example can be *I* for image, *T* for text, or *S* for structured tabular records

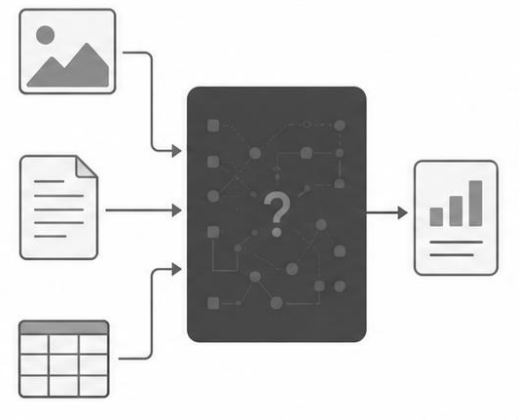
Takeaway CMC needs coherence at the class level, not forced equality at the instance level.

The pieces exist; CMC needs the composition

Black-box fusion can be strong; prototypes can explain; CMC needs an explainable model for complementary, missing, asymmetric evidence.

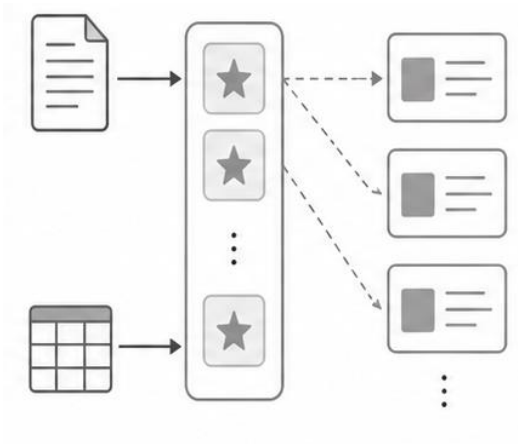
Black-Box Models

Large AutoML / transformer fusion



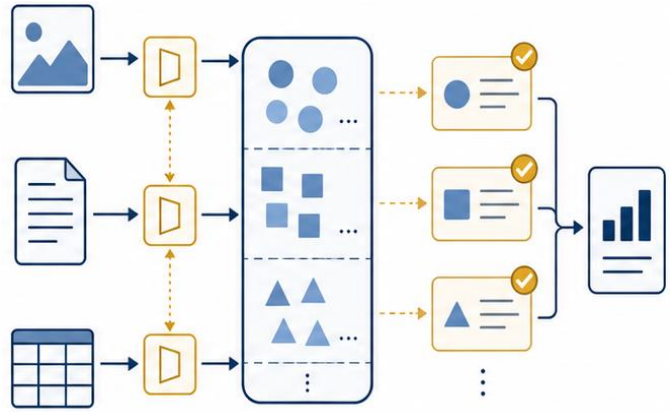
- Strong predictive performance
- Black-Box Predictions
- Not specific for CMC views
- Tends to overfit in low-data scenarios

SOTA Prototype networks



- Lower predictive performance
- Intrinsic case-based reasoning
- Usually unimodal or redundant-view
- Brittle under CMC degeneracies

CAMP (Ours)



- Same or Better prediction than Black-box
- Intrinsic case-based reasoning
- Class-wise prototype usage alignment
- Modality-agnostic inference

Takeaway There is a clear gap for explainable-by-design CMC without sacrificing performance.

RESEARCH QUESTIONS

Can prototypes stay accurate, coherent, and faithful in CMC?

Yes, CAMP achieves this through gated fusion, a partitioned prototype bank, and optimal transport losses that coherently align multimodal evidence

01 Performance

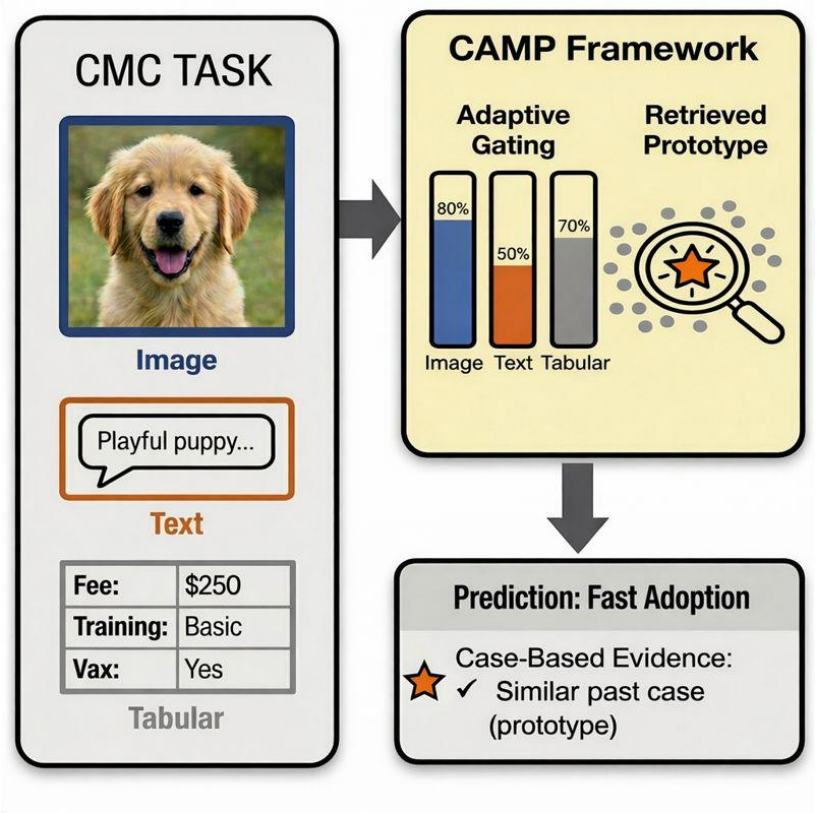
Can a compact prototype head match strong fusion baselines under fixed encoders?

02 Coherence

Can modalities stay complementary per instance while sharing class-level semantics?

03 Explainability

Can the same forward pass enable case-based reasoning explanations by desing?

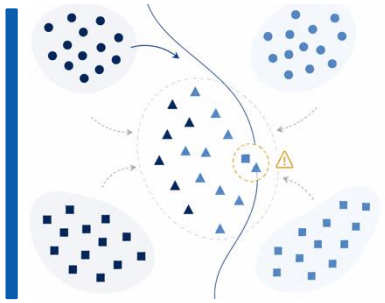


Our Method: • Align evidence usage — not individual instances • Class-Partitioned Prototype Bank • Expose gates, retrieved cases, and local evidence

CORE CHALLENGES

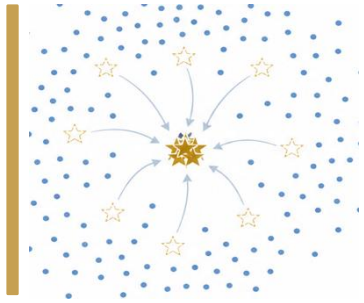
Complementarity amplifies prototype failure modes

A prototype bank for CMC must be separated, used, stable, and cross-modally coherent.



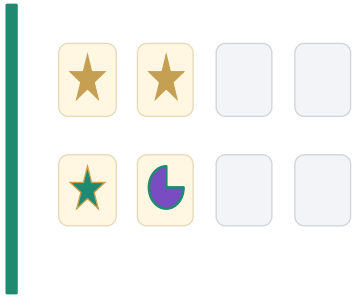
1 · Dominance / disagreement

One modality monopolizes geometry; modalities cluster classes differently



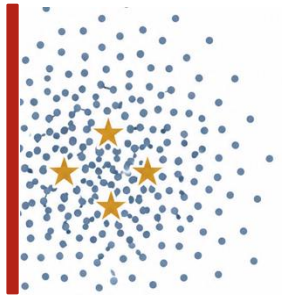
2 · Collapse

Several prototypes converge to the same local support



3 · Dead or impure units

Unused prototypes or mixed-class anchors dilute explanations



4 · Outlier drift

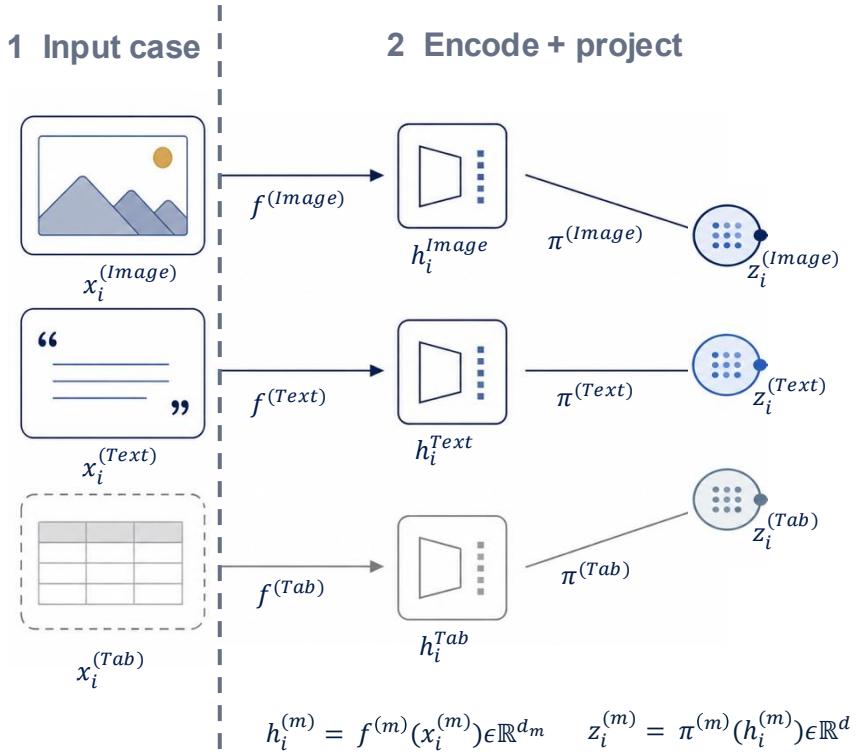
Rare multimodal cases pull prototypes away from support

Takeaway We should not merely add prototypes to a fusion model. We should make the prototype bank coherent and well-conditioned.

METHOD

CAMP inference: encode, gate, retrieve, explain

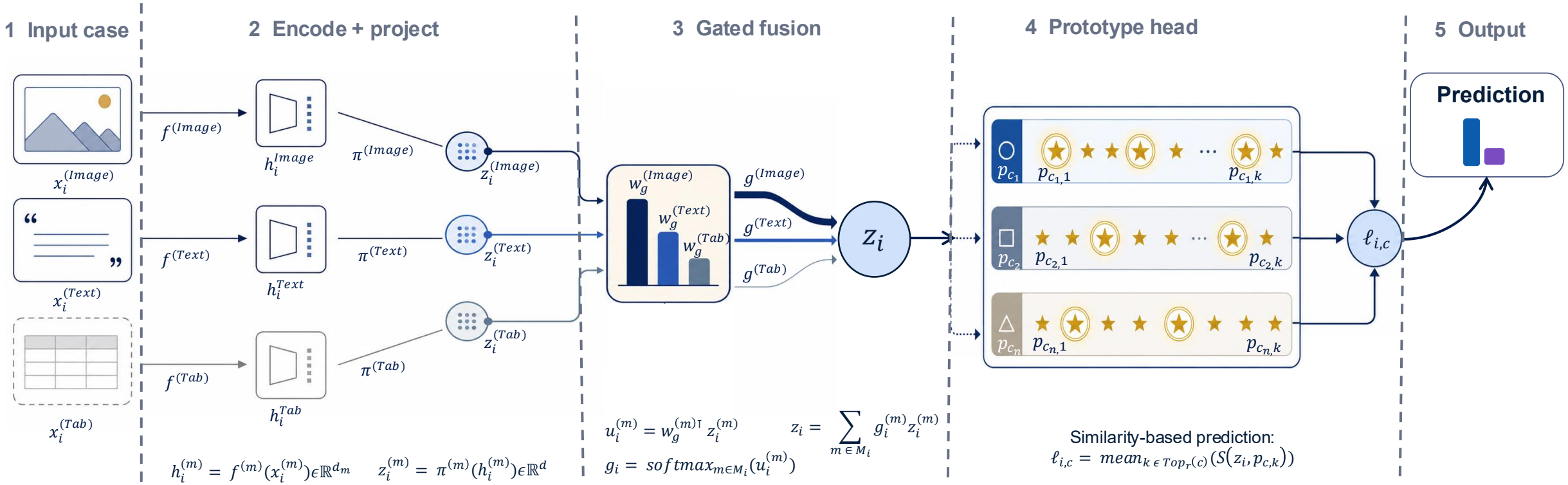
One model handles any observed subset of image, text, and tabular modalities; explanations are returned by the same forward pass.



METHOD

CAMP inference: encode, gate, retrieve, explain

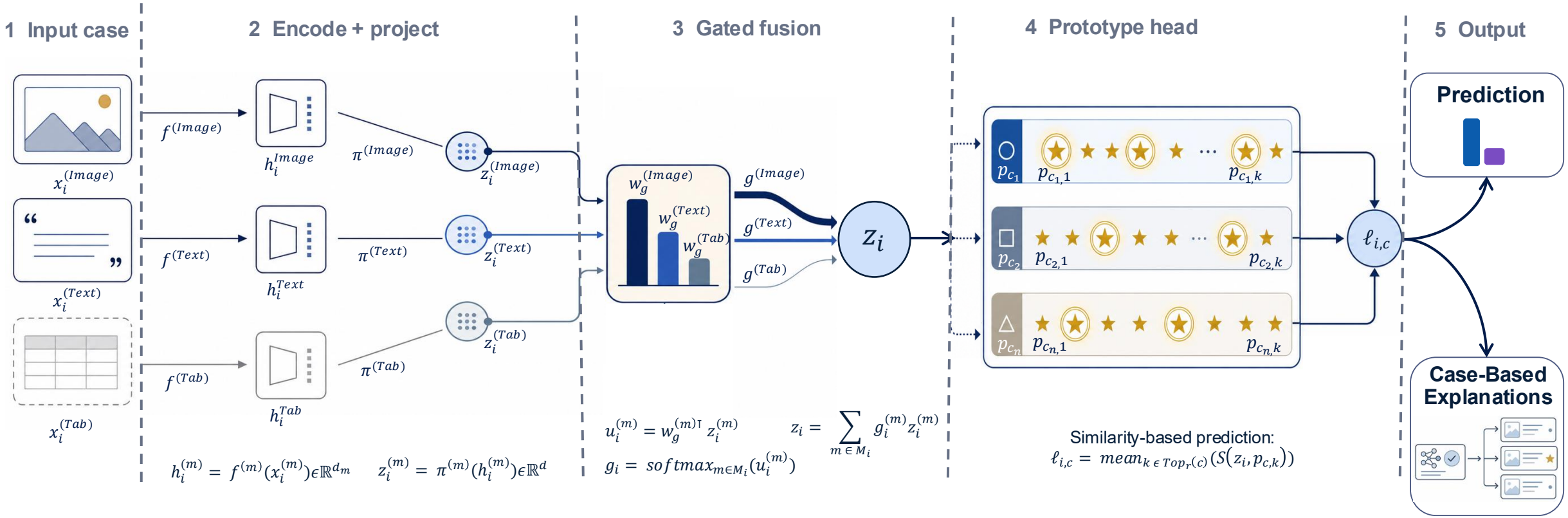
One model handles any observed subset of image, text, and tabular modalities; explanations are returned by the same forward pass.



METHOD

CAMP inference: encode, gate, retrieve, explain

One model handles any observed subset of image, text, and tabular modalities; explanations are returned by the same forward pass.

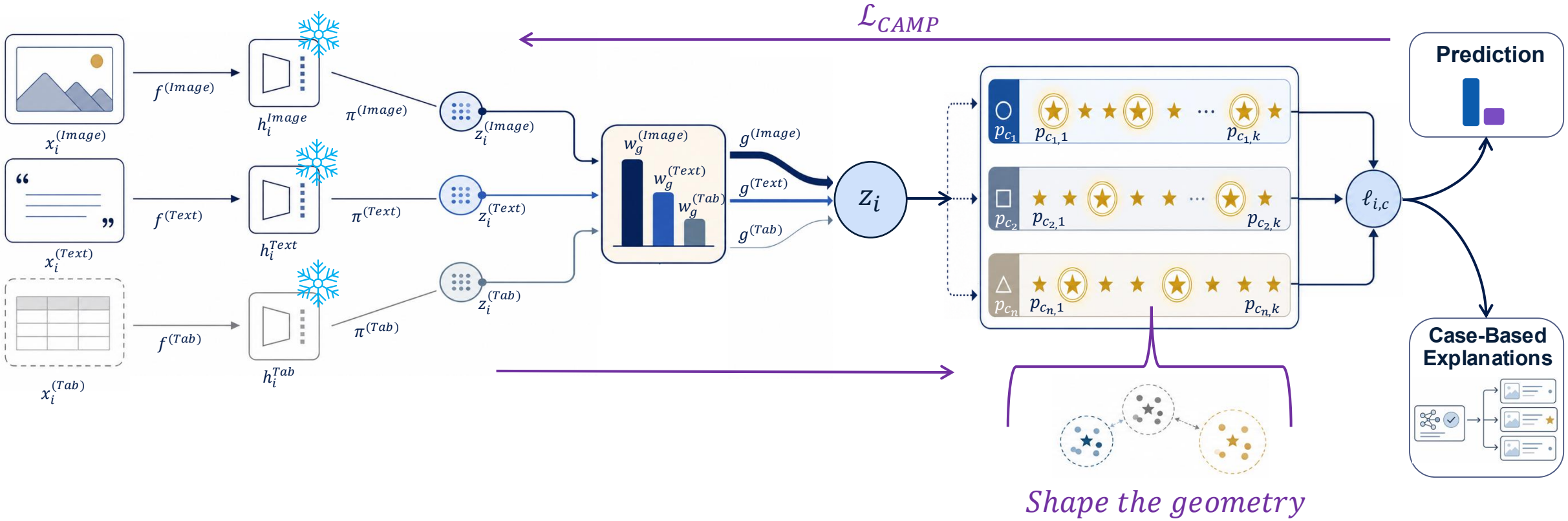


Takeaway A single forward pass produces both prediction and case-based explanations by design.

METHOD

CAMP inference: encode, gate, retrieve, explain

One model handles any observed subset of image, text, and tabular modalities; explanations are returned by the same forward pass.



Takeaway A single forward pass produces both prediction and case-based explanations by design.

Align evidence usage — not individual instances

01 Performance

Can a compact prototype head match large AutoML baselines?

02 Coherence

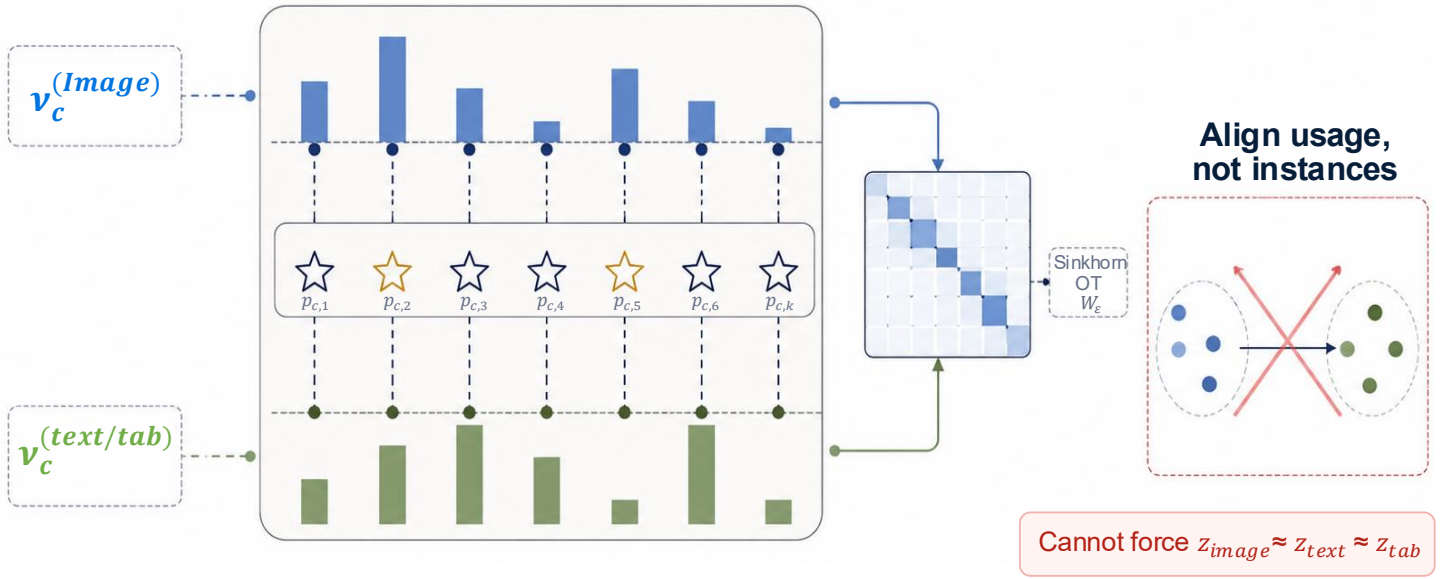
Can modalities stay complementary while sharing class semantics?

03 Explainability

Can the same forward pass enable case-based reasoning explanations by desing?

Class-wise prototype usage distributions

One class c: modalities place mass on the same prototype support



Key Takeaway Do not force modalities of the same case to coincide. Make them agree on which prototypes define a class.

Aligning Class Evidence to Match Black-Box Performance

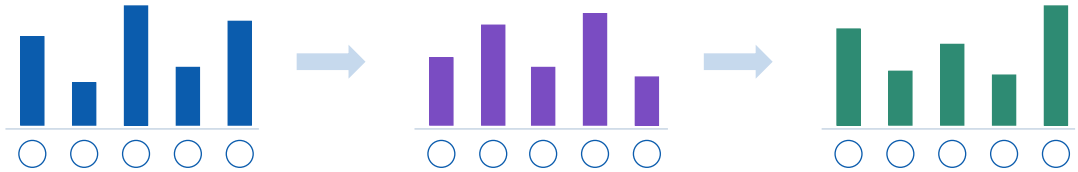
OT synchronizes modality-specific usage distributions, supported by targeted losses that eliminate prototype degeneracies.

$\mathcal{L}_{mm}$: Class-wise OT — Aligns usage across modalities

Controls: dominance / disagreement

For one class c, each modality induces a measure on the same prototypes

Class-wise prototype-usage measure



Multimodal class-wise OT loss → **Sinkhorn OT over prototype mass**

$\mathcal{L}_{geo}$: margins + diversity — separates classes; prevents collapse

controls: collapse

Full training objective

$$\mathcal{L}_{CAMP} = \mathcal{L}_{pred} + \lambda_{mm}\mathcal{L}_{mm} + \lambda_{gep}\mathcal{L}_{geo} + \lambda_{alloc}\mathcal{L}_{alloc} + \lambda_{anchor}\mathcal{L}_{anchor}$$

Note CAMP's theoretical contribution is the composition: geometry, allocation, anchors, and class-wise OT turn a prototype bank into a coherent object for prediction, missing-modality inference, and case-based explanation. * Equations and guarantees are in the paper; these losses are training-time only.

Aligning Class Evidence to Match Black-Box Performance

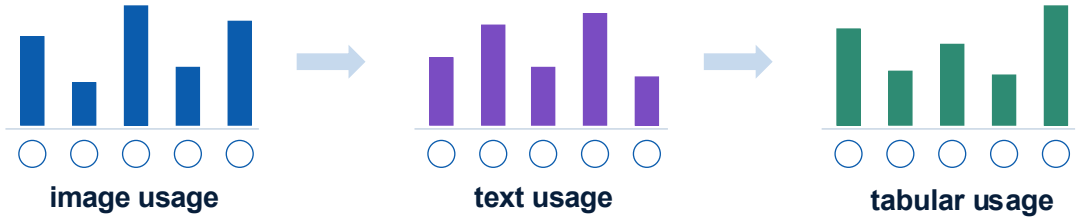
OT synchronizes modality-specific usage distributions, supported by targeted losses that eliminate prototype degeneracies.

$\mathcal{L}_{mm}$: Class-wise OT — Aligns usage across modalities

Controls: dominance / disagreement

For one class c, each modality induces a measure on the same prototypes

Class-wise prototype-usage measure



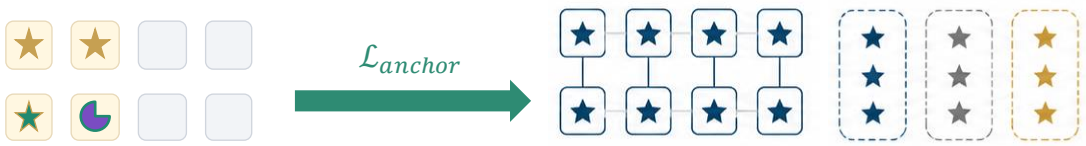
Multimodal class-wise OT loss → **Sinkhorn OT over prototype mass**

$\mathcal{L}_{geo}$: margins + diversity — separates classes; prevents collapse

controls: collapse

$\mathcal{L}_{alloc}$: coverage + purity — keeps prototypes used and class-specialized

controls: dead / impure units



Full training objective

$$\mathcal{L}_{CAMP} = \mathcal{L}_{pred} + \lambda_{mm}\mathcal{L}_{mm} + \lambda_{gep}\mathcal{L}_{geo} + \lambda_{alloc}\mathcal{L}_{alloc} + \lambda_{anchor}\mathcal{L}_{anchor}$$

Note CAMP's theoretical contribution is the composition: geometry, allocation, anchors, and class-wise OT turn a prototype bank into a coherent object for prediction, missing-modality inference, and case-based explanation. * Equations and guarantees are in the paper; these losses are training-time only.

Aligning Class Evidence to Match Black-Box Performance

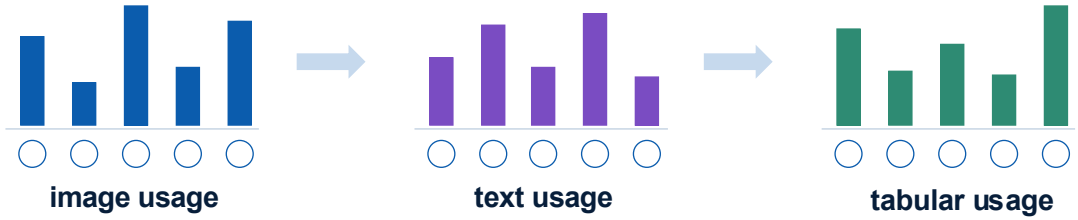
OT synchronizes modality-specific usage distributions, supported by targeted losses that eliminate prototype degeneracies.

$\mathcal{L}_{mm}$: Class-wise OT — Aligns usage across modalities

Controls: dominance / disagreement

For one class c, each modality induces a measure on the same prototypes

Class-wise prototype-usage measure



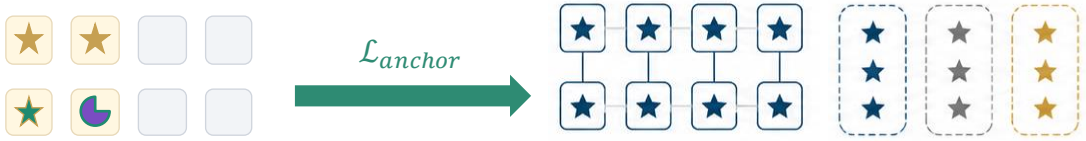
Multimodal class-wise OT loss

Sinkhorn OT over prototype mass

$\mathcal{L}_{alloc}$: coverage + purity

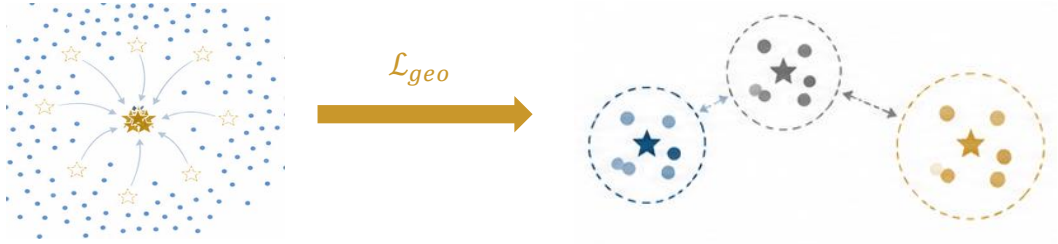
controls: dead / impure units

keeps prototypes used and class-specialized



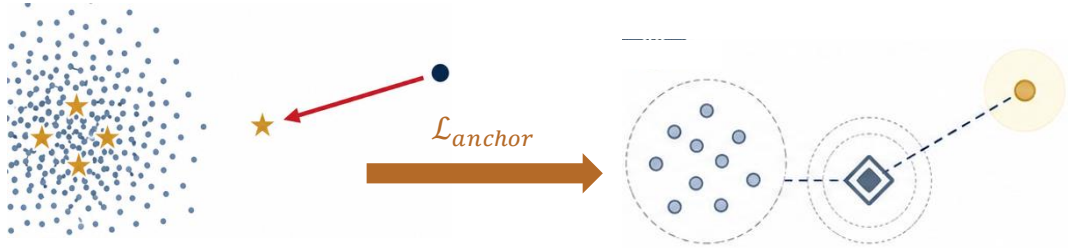
$\mathcal{L}_{geo}$: margins + diversity — separates classes; prevents collapse

controls: collapse



$\mathcal{L}_{anchor}$: rare-case anchors — stabilizes outliers and drift

controls: drift



Full training objective

$$\mathcal{L}_{CAMP} = \mathcal{L}_{pred} + \lambda_{mm}\mathcal{L}_{mm} + \lambda_{gep}\mathcal{L}_{geo} + \lambda_{alloc}\mathcal{L}_{alloc} + \lambda_{anchor}\mathcal{L}_{anchor}$$

Note CAMP's theoretical contribution is the composition: geometry, allocation, anchors, and class-wise OT turn a prototype bank into a coherent object for prediction, missing-modality inference, and case-based explanation. * Equations and guarantees are in the paper; these losses are training-time only.

EVALUATION PROTOCOL

16 public CMC datasets, controlled comparisons

The benchmark separates encoder capacity from the prototype objective.

Dataset map

Tabular + Text
6 datasets
IMDB · FakeJob · Kick · Jigsaw · Wine · Airbnb

Image + Text
4 datasets
PTech · Food101 · Fakeddit · Memotion

Image + Tabular
3 datasets
CCD · WikiArt · HAM

Image + Text + Tabular
3 datasets
PetFinder · COVID · Artm

Experimental Evaluation

CAMP-F — Frozen Swin-L + DeBERTa-v3 embeddings; ~660K trainable head

CAMP-E — End-to-end fine-tuning after head-only warmup

Baselines — XGB, LP, Proto-Raw, Proto-Base, ProtoMedX, MLP, TTT, AutoGluon

Metrics — AUC/Acc/QWK; mean ± std over five seeds; Avg and MRR

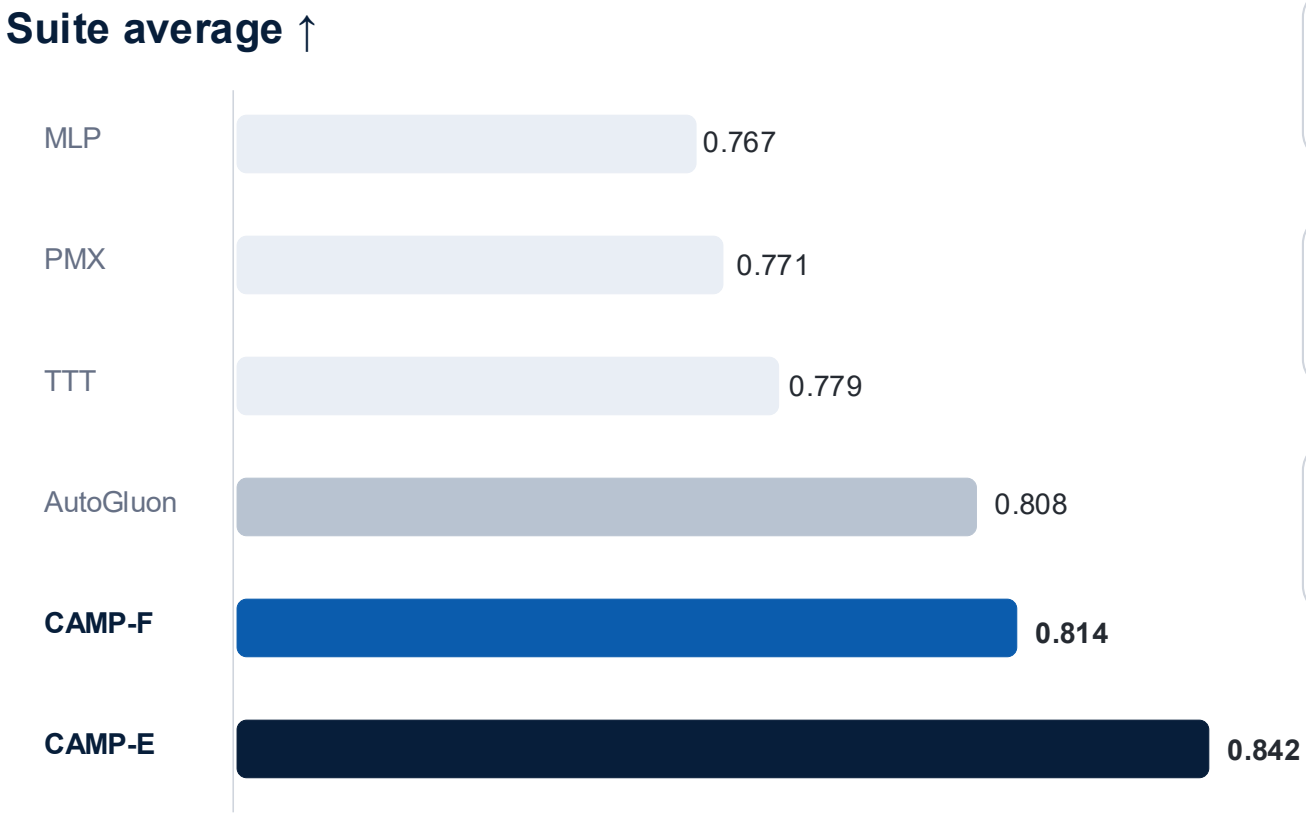
Tuning — 50 Optuna trials per dataset

Note Frozen comparisons use identical embeddings to isolate the head.

RESULTS: QUANTITATIVE

Compact head, large-scale performance

CAMP-F isolates the structural objective; CAMP-E shows the ceiling with adaptation.



0.814
CAMP-F suite average

0.842
CAMP-E suite average

11/16
CAMP-F ≥ AutoGluon

16/16
CAMP-E best

0.448
CAMP-F MRR vs AG (0.296)

1.000
CAMP-E MRR

Key Takeaway: Frozen-encoder CAMP-E uses the same fixed embeddings as the MLP yet improves Avg 0.767 → 0.814 and MRR 0.174 → 0.448.

Mechanistic validation: each block fixes its failure mode

The ablation validates the theory: OT helps most after support is separated, allocated, and anchored.

Cumulative Ablations of Loss Components.

Proto-Classic	0.7445
+ Geometry block	0.7538
+ Allocation block	0.7658
+ Anchor block	0.7852
+ Class-wise OT	0.8139

Objectives cause the intended effect:

Similarity margin

0.031 → 0.118

Class separation expands

Dead prototypes

26.6% → 2.3%

Unused Prototypes become active

Prototype impurity

0.22 → 0.04

Class specialization improves

Cross-modal Wε

0.40 → 0.12

Usage disagreement drops

Key Takeaway: Proto-Classic 0.7445 → CAMP full 0.8139 (+0.0694) Largest leave-one-out hit: remove OT = −0.029

A single model degrades gracefully and explains its evidence

CAMP uses only observed modalities at inference and returns model-internal evidence.

Missing-modality robustness

No imputation or retraining

Single-drop avg	**CAMP-E 0.724**	AG 0.692
All configurations	**CAMP-E 0.634**	AG 0.610
HAM image-drop	**CAMP-F 0.894**	AG 0.493
PetFinder tab-drop	**CAMP-F 0.425**	AG 0.301

Key Takeaway: CAMP is able to work under partially observed streams.

A single model degrades gracefully and explains its evidence

CAMP uses only observed modalities at inference and returns model-internal evidence.

Missing-modality robustness

No imputation or retraining

Single-drop avg	**CAMP-E 0.724**	AG 0.692
All configurations	**CAMP-E 0.634**	AG 0.610
HAM image-drop	**CAMP-F 0.894**	AG 0.493
PetFinder tab-drop	**CAMP-F 0.425**	AG 0.301

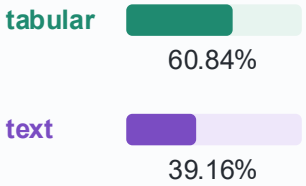
Qualitative case study of CAMP's Multimodal explainability

Not a post-hoc surrogate: model-internal fixed-support probe reuses CAMP gates, distances, and nearest prototypes.

Airbnb case 744: Predicted Class 5 ($100–$120) · Confidence 98.34%

Gate Values

Observed streams only



Nearest Prototypes

- ★ Class 5 proto. 1
- ★ Class 5 proto. 2
- ★ Class 5 proto. 3

Fixed-support Attribution

- \+ num_bedrooms
- \+ num_bathrooms
- \+ "prime location"
- − "Opposite of Sthybus station"

Key Takeaway: CAMP is able to work under partially observed streams.
Attribution is performed over the model's case-based reasoning with faithfulness guarantees (See Appendix)

TAKEAWAYS

Interpretable CMC does not require sacrificing performance

01 Formalizes CMC

Modalities complete rather than re-describe one another

02 Coherent Alignment: Aligns usage, not instances

Class-wise OT matches prototype usage distributions

03 Explains from the model itself

Gates + retrieved cases + fixed-support evidence

04 Structural bias over scale

Sub-1M frozen head rivals >100M black-box systems; E2E improves further

Best used as a drop-in CMC head when frozen encoders, auditability, and missing-modality robustness matter.

Thank you.