DRA: STATE-CONDITIONED EVIDENCE ALLOCATION FOR MISSING-MODALITY FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

When a modality disappears, the value of evidence still available but unread can change. We study *state-conditioned evidence allocation*: using observed modality availability to direct a limited budget of additional text encoding. A Bayes-risk characterization relates this choice to information substitutes and complements: the preferred state can reverse with the loss, and at an interior budget, state-blind allocation is optimal only at equal value per cost. DRA (Decide, Read, Adapt) implements the allocation with a state-gated residual reader that reuses an adapted CLIP text tower; we retain shared adaptation and validation thresholds as interfaces. On MM-IMDb, same-movie image removal with symmetric reader training gives a positive log-loss interaction of 0.00839 nats per label (movie-cluster 95% CI $[0.00767, 0.00905]$). With the reader and extra-window budget fixed, targeting image-missing rows improves macro-F1 by $+0.18$ and AP by $+0.33$ over state-blind allocation. In retrained controls, DRA comes within 0.19 F1 of all-row reading using 53% of the extra windows. Under published evaluation rules across MM-IMDb, UPMC-Food101 and Hateful Memes, DRA exceeds reported MoRA in 23 of 27 benchmark settings.

1 INTRODUCTION

Multimodal classifiers must often predict with an absent modality: a movie without a poster, a meme without extracted text, or a product without an image. With pretrained vision–language backbones such as CLIP (Radford et al., 2021), missing-aware prompts condition parameters on availability (Lee et al., 2023; Jang et al., 2024; Hu et al., 2024), low-rank methods couple encoders or allocate adapters to modality combinations (Zhao et al., 2026; 2025), and TLP changes missing-case logits (Chen & Guo, 2026). These interventions concern parameters and predictions. We study an additional choice at the input: *where should a limited encoding budget be spent when a modality is absent?* A fixed CLIP window reads only 75 content tokens; 62% of first-plot MM-IMDb summaries exceed it. A classifier can therefore lose its image while discarding part of the text it still possesses.

Our central observation is that *the value of the same unread evidence depends on which other modality is available*. A tail that repeats label evidence supplied by an image becomes more useful when that image disappears. We formalize this as *state-conditioned evidence allocation*: modality availability guides additional reading under a fixed budget. Bayes-risk analysis connects this choice to information substitutes and complements (Börgers et al., 2013): even conditionally independent views can favor opposite reading priorities under log loss and classification error. Within state-level policies, allocating by value per cost is optimal, and state-blind allocation requires equal ratios at an interior budget. The gate uses the observed state before tail encoding, retaining the pretrained context length.

DRA (*Decide, Read, Adapt*; Section 4) implements this allocation through a normalized residual reader. Read routes tail windows to image-missing examples and injects their features through the same adapted text tower. Adapt supplies a shared/private additive residual across states; Decide fits validation thresholds to compare scores at a common metric-facing interface. The central intervention is the state-conditioned evidence path, supported by shared adaptation and metric-aware decisions. Treating data-independent missingness as information garbling (Blackwell, 1953) connects acquisition to evaluation: it re-prices unread evidence and can lower the attainable F1 operating

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.

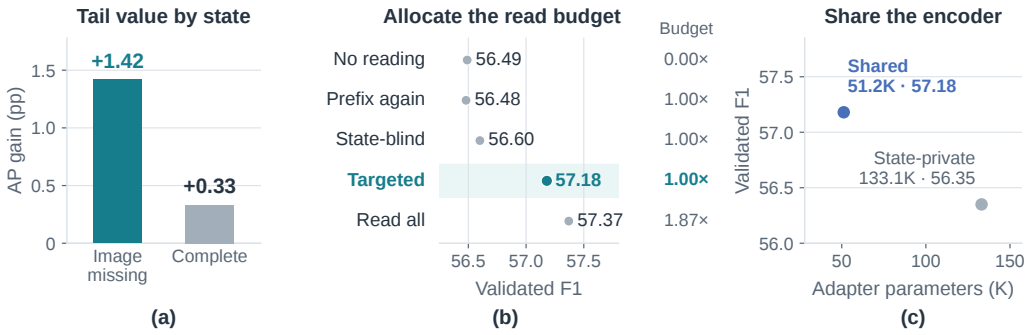


Figure 1: **Evidence allocation and its encoder interface.** MM-IMDb, both missing at 70%. (a) Three tail windows give a larger AP gain without the image ($K = 3$ versus $K = 0$). (b) At a matched $1.00\times$ extra-window budget, targeting image-missing rows exceeds prefix re-reading and state-blind allocation; reading every eligible row costs $1.87\times$. Scores are validated F1. (c) The shared encoder uses fewer trainable adapter parameters and has a higher mean than the tested state-private version. Controlled results appear in Table 3.

point (Lipton et al., 2014). We therefore quantify reading value with log loss and compare F1 under a common threshold-fitting procedure. Shared adaptation supplies the encoder interface, evaluated through parameter-sharing controls.

On MM-IMDb, targeting comes within 0.19 F1 of all-row reading at 53% of its extra windows. Under published rules, DRA exceeds reported MoRA in 23 of 27 settings across three benchmarks; matched decision rules further quantify ranking and operating-point effects. Our contributions are:

- **Missingness as an evidence-acquisition signal.** We characterize the marginal value of unread evidence for general losses, exhibit a reversal of reading priority across losses, and identify when state-blind allocation can be optimal.
- **A state-gated residual reader.** DRA realizes this allocation through a shared CLIP text tower, preserving its context length and routing extra windows without a separate policy network.
- **Controlled identification of the allocation effect.** Same-movie image removal and symmetric reader training test the reading-value interaction; fixed-reader, identical-budget comparisons isolate the allocation effect. Shared-encoder controls, matched decision rules and cross-scenario tests evaluate the supporting interfaces.

2 Related Work

Missing modalities. Beyond modality dropout, reconstruction and distillation (Neverova et al., 2016; Zhao et al., 2021; Wei et al., 2023; Wu et al., 2026), missing-aware prompts condition a frozen backbone on each missing case (Lee et al., 2023; Jang et al., 2024; Hu et al., 2024; Lang et al., 2025; Zhang et al., 2025); MoRA couples the CLIP towers through shared low-rank factors (Zhao et al., 2026), and MCULoRA allocates adapters to missing-modality combinations (Zhao et al., 2025). Adapt retains MoRA's cross-modal coupling in an additive residual. DPL (Lu et al., 2026) learns state-specific, modality-decomposed class prototypes on CLIP; TLP (Chen & Guo, 2026) learns test-time logit offsets. DRA additionally allocates encoder passes to unread observed text and validates its decision thresholds without learning state-specific prototypes.

Reading more of an observed modality. Texts beyond the 77-token window of CLIP (Radford et al., 2021) are truncated (Sun et al., 2019), encoded in chunks (Pappagari et al., 2019) or handled by sparse attention (Beltagy et al., 2020) and stretched positional embeddings (Zhang et al., 2024a). Adaptive computation (Graves, 2016; Schwartz et al., 2020), skimming (Yu et al., 2017) and feature acquisition (Xu et al., 2012; Janisch et al., 2019) decide how much input to process, and SIEVE learns whether to repair a missing modality (Gao et al., 2026). Information substitutes and

complements describe how one signal changes another's decision value (Börgers et al., 2013). We apply that view to missingness: which state should receive an observed but unread tail under a fixed budget? DRA's state-level gate is available before reading, retains the backbone's context length, and is evaluated with a fixed reader and matched budgets.

Decision rules. F1-optimal classifiers threshold calibrated probabilities at metric-dependent levels (Koyejo et al., 2014; Narasimhan et al., 2014): for binary F1, half the optimal F1 is an optimal threshold (Lipton et al., 2014), and macro-averaged F1 decomposes into label-wise thresholds (Koyejo et al., 2015; Pillai et al., 2013). Example weighting changes the effective cost ratio and hence the threshold (Zadrozny et al., 2003); calibration (Platt, 2000; Guo et al., 2017) and prior correction (Saerens et al., 2002; Lipton et al., 2018) change decisions without changing the representation. The MoRA benchmark nevertheless scores MM-IMDb macro-F1 at a fixed threshold of 0.5 (Zhao et al., 2026).

3 State-conditioned evidence allocation

Observation model. Let $X = (V, P, U)$ contain an image V, the text prefix P and the unread tail U. For each binary label Y, $S \in \{\mathrm{F}, \mathrm{T}, \mathrm{I}\}$ records both modalities, text only or image only. Under data-independent permanent missingness, $S \perp (X, Y)$ and the prefix-only observation is

$$O = (S, h_S(X)), \qquad h_{\mathrm{F}}(X) = (V, P), \quad h_{\mathrm{T}}(X) = P, \quad h_{\mathrm{I}}(X) = V. \tag{1}$$

A scenario specifies the state distribution. Missingness is a Blackwell garbling of (V, P) (Blackwell, 1953), with two consequences: it changes the value of acquiring U and can lower the attainable F1 operating point. Although $I(Y; S) = 0$, the state can still determine which evidence is worth reading.

3.1 Missingness re-prices unread evidence

Condition on a valid tail $U = U_{\le K}$ and consider $s \in \{\mathrm{T}, \mathrm{F}\}$. For a loss ℓ, write $H_\ell(Y \mid Z) = \inf_f \mathbb{E}[\ell(Y, f(Z))]$ for its finite Bayes risk and $I_\ell(Y; U \mid Z) = H_\ell(Y \mid Z) - H_\ell(Y \mid Z, U)$ for the value of reading U. Set $\Delta_{\mathrm{T}} = I_\ell(Y; U \mid P)$ and $\Delta_{\mathrm{F}} = I_\ell(Y; U \mid V, P)$. For state probabilities p_s, expected costs $c_s > 0$ and reading probabilities ρ_s, the allocation problem is

$$\max_{0 \le \rho_s \le 1} \sum_s p_s \rho_s \Delta_s \qquad \text{s.t.} \qquad \sum_s p_s \rho_s c_s \le B. \tag{2}$$

Selection is independent of the example within a state; reading is observed by the predictor. Log loss gives $I_\ell = I$, the usual conditional mutual information.

Proposition 1 (Missingness re-prices unread evidence)**.** ***(a)*** *For any loss with finite Bayes risks,*

$$\Delta_{\mathrm{T}} - \Delta_{\mathrm{F}} = I_\ell(Y; V \mid P) - I_\ell(Y; V \mid U, P). \tag{3}$$

The tail is worth more without the image exactly when the image is worth more without the tail: the views are strict substitutes for this decision problem (Börgers et al., 2013). ***(b)*** *For log loss, the difference is $I(U; V \mid P) - I(U; V \mid P, Y) \ge 0$ whenever $U \perp V \mid (P, Y)$, assuming finite information terms. Under 0–1 loss, such conditionally independent views can instead be complements.* ***(c)*** *Eq. (2) is solved by reading states in decreasing Δ_s / c_s. If $p_{\mathrm{T}}, p_{\mathrm{F}} > 0$ and $0 < B < \sum_s p_s c_s$, state-blind allocation ($\rho_{\mathrm{T}} = \rho_{\mathrm{F}}$) can be optimal only if $\Delta_{\mathrm{T}}/c_{\mathrm{T}} = \Delta_{\mathrm{F}}/c_{\mathrm{F}}$.*

Proof sketch. Expanding the four Bayes risks gives (a); the log-loss chain rule gives (b). For its reversal, let $\Pr(Y = 1) = 0.8$ and let U, V independently equal Y with probability 0.7, conditional on Y. Under 0–1 loss, $\Delta_{\mathrm{T}} = 0$ but $\Delta_{\mathrm{F}} = 0.026$; under log loss, $\Delta_{\mathrm{T}} - \Delta_{\mathrm{F}} = I(U; V) > 0$. For (c), shifting cost to a higher gain/cost ratio strictly improves the objective (Appendix C).

Part (b) connects the conditional-independence assumption used in co-training (Blum & Mitchell, 1998) to information substitutes, but shows that the ordering depends on the loss. We quantify reading value with log loss and assess F1 under matched decision rules. For equal costs and $\Delta_{\mathrm{T}} > \Delta_{\mathrm{F}}$, $B = p_{\mathrm{T}} c_{\mathrm{T}}$ gives $\rho_{\mathrm{T}} = 1, \rho_{\mathrm{F}} = 0$: DRA's gate solves this state-level problem without a policy network. With M modalities, the same allocation rule applies to up to $2^M - 1$ nonempty availability states, given one candidate read and cost per state; we instantiate three image–text states.

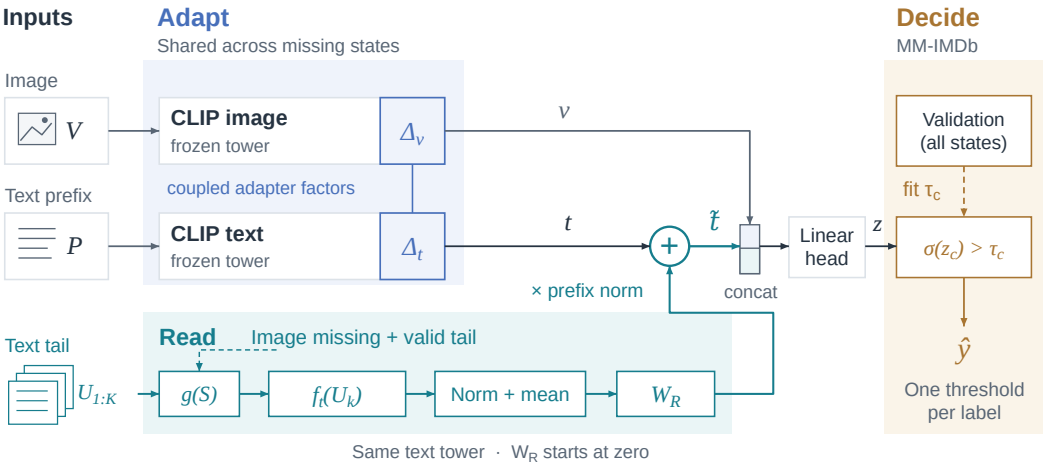


Figure 2: **DRA overview.** The missing state gates tail reading through a shared adapted text tower. Adapt shares encoder parameters across states; Decide fits pooled per-label validation thresholds on MM-IMDb.

We test **(R1)** the sign using the same movies with and without the image, including symmetric reader training; and **(R2)** targeted versus state-blind allocation with the reader and budget fixed. Prefix re-reading adds no Bayes information and controls for computation. These are tests with fitted predictors; under MCAR the Bayes Δ_s does not depend on the missing ratio, whereas learned-reader gains can.

3.2 Garbling and the evaluation interface

Let $\mu(O) = \Pr(Y = 1 \mid O)$ and $\pi = \Pr(Y = 1) > 0$. Population F1 is $F(A) = 2\Pr(A, Y = 1)/(\Pr(A) + \pi)$. A richer observation can simulate any garbled decision rule, so its optimal F1 cannot be lower (Lemma 1, Appendix C). By Lipton et al. (2014), a canonical optimal posterior threshold is $\tau^\star = F^\star(O)/2 \le 1/2$. Thus data-independent modality removal cannot raise either the optimum or this canonical cutoff. Pooled F1 uses one posterior cutoff across states; we use one pooled validation threshold per label to evaluate Read.

Class weighting also moves the operating point: unrestricted weighted-BCE minimization gives $z_w = \operatorname{logit}\mu + \log w$, the classical cost-sensitive result (Elkan, 2001; Zadrozny et al., 2003). It preserves ranking at the population optimum. We check **(D1)** calibrated thresholds against state-wise validation estimates $\hat{F}_s^\star/2$; **(D2)** the empirical hypothesis that threshold fitting helps missing states more; and **(D3)** logit shifts against $\log w$. These diagnostics motivate the common decision interface used in reading comparisons.

4 DRA: A state-conditioned evidence allocator

DRA implements state-conditioned reading through an image-missing gate (Figure 2). We retain shared adaptation and validation thresholding as its encoder and evaluation interfaces. We train the residual adapters, reader and linear head with binary cross-entropy for multilabel tasks and cross-entropy otherwise; CLIP is frozen. Missing inputs are white images and empty text, and the head concatenates the projected tower embeddings. Fixed-reader, fixed-budget controls isolate allocation from encoder retraining.

4.1 Read: state-gated residual evidence acquisition

We split text into a prefix P and up to K consecutive tail windows of 75 content tokens. For nonempty valid-tail set $\mathcal{K}$, the state gate is $g(S) = \mathbf{1}\{S = \mathrm{T}\}$. Using the same adapted text tower

for $t = f_{\mathrm{t}}(P)$ and the tail,

$$\tilde{t} = t + g(S)\|t\|W_R\bar{u}, \qquad \bar{u} = \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\frac{f_{\mathrm{t}}(U_k)}{\|f_{\mathrm{t}}(U_k)\|}. \tag{4}$$

Rows with no valid tail keep t; $g(S) = 0$ skips tail encoding. The zero-initialized $W_R \in \mathbb{R}^{d\times d}$ makes Read initially a no-op, and normalization plus $\|t\|$ match the prefix scale. Read adds $d^2 = 262{,}144$ parameters ($d = 512$), reuses the tower and needs no policy network. We fix $K = 3$, covering 98% of training-audited tokens with the prefix. Each triggered row adds up to three tower passes. Food101 and Hateful Memes use the prefix-only path. The gate implements the text-only-first ordering motivated above; Section 5.4 measures its benefit and cost.

4.2 Shared encoder interface (Adapt)

We share encoder parameters because availability changes the evidence to acquire, while each tower still represents the same observed modality; **(A1)** shared versus state-private controls test this design. Adapt retains MoRA's projections, blocks and ranks (Table 10). With row-vector projections $x \mapsto xW_m$, it adds

$$\Delta W_{\mathrm{t}} = \kappa_{\mathrm{t}}\big(A(B^\top B)A^\top + C_{\mathrm{t}}D_{\mathrm{t}}\big), \qquad \Delta W_{\mathrm{v}} = \kappa_{\mathrm{v}}\big(B(A^\top A)B^\top + C_{\mathrm{v}}D_{\mathrm{v}}\big), \tag{5}$$

where $A \in \mathbb{R}^{d_{\mathrm{t}}\times r}$, $B \in \mathbb{R}^{d_{\mathrm{v}}\times r}$, $C_m \in \mathbb{R}^{d_m\times r_m}$, $D_m \in \mathbb{R}^{r_m\times d_m}$ and $\kappa_m = \alpha/r_m$. The shared factors couple towers through the partner's Gram matrix. MoRA normalizes its shared/private weights using learnable magnitudes (Zhao et al., 2026; Liu et al., 2024); Adapt directly adds the combined residual. Both permit merged inference; the parameterization is the difference. Shared factors use $\mathcal{N}(0, 0.01^2)$ and private right factors are zero-initialized. The adapter is state-agnostic; availability gates evidence acquisition through the common encoder.

4.3 Evaluation interface (Decide)

For MM-IMDb, each label uses the threshold maximizing validation F1 in the matching scenario, pooling states. Candidates are score-order cuts including all-negative/all-positive endpoints; ties choose the larger threshold. Validation and test are disjoint. Strictly increasing score transforms preserve the selected cuts and test predictions, whereas biases can change fixed-threshold F1. Both models use the same fitting procedure with separately fitted thresholds. Thus Decide exposes the operating point when assessing evidence acquisition. Accuracy and AUROC require no fitted threshold, so Food101 and Hateful Memes evaluate Adapt alone.

5 Experiments

5.1 Setup

We follow the permanent-missingness protocol of Zhao et al. (2026) with CLIP ViT-B/16 (Radford et al., 2021) on MM-IMDb (Arevalo et al., 2017) (23-label macro-F1), UPMC-Food101 (Wang et al., 2015) (101-way accuracy) and Hateful Memes (Kiela et al., 2020) (binary AUROC). At missing ratios $\eta \in \{50\%, 70\%, 90\%\}$, η of examples lack the image, lack text, or are split equally between these two states (the both-missing scenario); the remainder are complete. Released tables fix these data-independent assignments in every split. MM-IMDb scores one row per plot, with movie-shared poster, labels and state. An unavailable image is white and unavailable text is empty; hidden modalities are never used in training. One model is trained per ratio and scenario, with the same state distribution for checkpoint selection and test; split counts are in Appendix B.

We compare with published CoOp (Zhou et al., 2022), MMP (Lee et al., 2023), MaPLe (Khattak et al., 2023), DePT (Zhang et al., 2024b), DCP (Hu et al., 2024) and MoRA (Zhao et al., 2026), and use MoRA† (official code and configuration) when model outputs are needed. DRA uses Adapt on all datasets and Read (K=3) on MM-IMDb, with one recipe per dataset (Table 10). Recipes are selected at η=70% using the stated validation metrics; checkpoint selection uses the official metric. Table 1 uses the official MM-IMDb rule (sigmoid > 0.5), with three-seed means. Reading and encoder ablations use both missing at η=70% with up to six seeds. MoRA† retains its official adapter,

Table 1: **Comparison with published results** (CLIP ViT-B/16, permanent-missingness protocol). Image/Text columns denote the proportions observed (the complement is missing). CoOp–MoRA: values reported by Zhao et al. (2026); DRA: mean of three seeds, with Δ to reported MoRA. MM-IMDb uses the official rule (sigmoid > 0.5). Bold: best per row.

Dataset (metric)	η	Image	Text	Reported in Zhao et al. (2026) CoOp	MMP	MaPLe	DePT	DCP	MoRA	DRA (Δ)
MM-IMDb (F1-Macro)	50%	100%	50%	48.06	48.88	49.58	50.64	52.13	54.62	**55.41 (+0.79)**
		50%	100%	49.89	51.46	52.32	52.78	54.32	57.61	**57.96 (+0.35)**
		75%	75%	48.37	49.32	49.56	50.87	52.32	55.88	**56.23 (+0.35)**
		Average		48.77	49.89	50.49	51.43	52.92	56.04	**56.53 (+0.49)**
	70%	100%	30%	44.13	45.64	45.52	46.38	48.52	52.56	**52.72 (+0.16)**
		30%	100%	48.82	50.52	50.64	52.13	53.14	56.39	**56.81 (+0.42)**
		65%	65%	46.84	48.12	49.16	50.32	51.42	52.97	**53.35 (+0.38)**
		Average		46.60	48.09	48.44	49.61	51.03	53.97	**54.30 (+0.33)**
	90%	100%	10%	44.76	45.32	46.84	47.56	49.26	50.67	**50.68 (+0.01)**
		10%	100%	48.32	49.12	50.13	50.88	52.22	53.57	**55.70 (+2.13)**
		55%	55%	44.12	44.87	45.12	46.54	48.04	51.64	**51.87 (+0.23)**
		Average		45.73	46.44	47.36	48.33	49.84	51.96	**52.75 (+0.79)**
UPMC-Food101 (Accuracy)	50%	100%	50%	77.45	77.89	79.64	80.16	82.11	84.41	**85.20 (+0.79)**
		50%	100%	87.02	87.16	87.35	82.14	89.12	89.63	**89.83 (+0.20)**
		75%	75%	81.24	81.72	82.34	83.12	85.24	86.68	**87.24 (+0.56)**
		Average		81.90	82.26	83.11	81.81	85.49	86.91	**87.42 (+0.51)**
	70%	100%	30%	76.34	76.52	77.02	77.34	78.87	80.85	**81.76 (+0.91)**
		30%	100%	84.78	85.64	85.89	86.12	87.32	88.01	**88.31 (+0.30)**
		65%	65%	78.87	79.12	79.84	81.46	81.87	83.77	**84.62 (+0.85)**
		Average		80.00	80.43	80.92	81.64	82.69	84.21	**84.90 (+0.69)**
	90%	100%	10%	71.87	73.14	73.46	74.12	75.26	78.41	**79.12 (+0.71)**
		10%	100%	81.67	82.14	83.12	83.56	85.78	86.77	**86.86 (+0.09)**
		55%	55%	76.46	76.58	77.85	78.12	79.87	81.09	**81.61 (+0.52)**
		Average		76.67	77.29	78.14	78.60	80.30	82.09	**82.53 (+0.44)**
Hateful Memes (AUROC)	50%	100%	50%	60.56	60.31	60.87	61.87	62.32	70.66	**72.80 (+2.14)**
		50%	100%	62.41	62.35	63.13	63.88	64.46	**71.58**	71.40 (-0.18)
		75%	75%	64.87	65.84	65.46	65.86	66.02	69.58	**70.06 (+0.48)**
		Average		62.61	62.83	63.15	63.87	64.27	70.61	**71.42 (+0.81)**
	70%	100%	30%	60.74	61.12	61.26	61.56	62.82	69.43	**71.46 (+2.03)**
		30%	100%	62.74	63.24	63.14	63.48	64.12	**70.68**	69.14 (-1.54)
		65%	65%	64.82	65.04	65.23	65.48	66.08	**70.15**	69.71 (-0.44)
		Average		62.77	63.13	63.21	63.51	64.34	70.09	**70.10 (+0.01)**
	90%	100%	10%	60.03	57.21	60.74	61.14	62.08	68.52	**70.76 (+2.24)**
		10%	100%	61.46	61.52	61.87	62.42	63.87	**68.78**	68.06 (-0.72)
		55%	55%	64.32	63.34	64.85	65.37	66.78	68.37	**68.78 (+0.41)**
		Average		61.94	60.69	62.49	62.98	64.24	68.56	**69.20 (+0.64)**

optimiser and loss recipe while sharing masks, splits, seeds, evaluation and checkpoint-selection protocol with the corresponding DRA controls; numerical hyperparameters remain method-specific. Tables report seed means; paired differences use shared seeds.

5.2 Comparison with published results

Under the published evaluation rules, DRA exceeds reported MoRA in 23 of 27 settings, including all nine on MM-IMDb and UPMC-Food101 (Table 1). Mean margins are $+0.54$, $+0.55$ and $+0.49$ across the three datasets. This count summarizes the released benchmark settings; the paired seed controls below quantify the variation directly. MoRA† differs from reported values by -0.43, -0.29 and -0.44 on average (Appendix A).

On Hateful Memes, the largest variation occurs in image-missing, text-rich settings, while both-missing averages are close to MoRA. Since its texts fit the prefix, these rows isolate the state-agnostic Adapt interface and complement the targeted-reading analysis on MM-IMDb.

Table 2: **Decide: comparing models under the same decision rule.** MM-IMDb test F1-Macro, mean over three seeds. Each block matches the loss and scenario; columns change only the decision rule. Δ: paired DRA$-$MoRA†; †: official-code reproduction. Bold: better mean within each pair, including ties.

Model	Fixed threshold 0.5			Validation thresholds		
	η=50%	η=70%	η=90%	η=50%	η=70%	η=90%
(a) Three-scenario average; unweighted loss						
MoRA†	55.01	53.39	52.29	57.21	55.89	54.59
DRA	**56.53**	**54.30**	**52.75**	**59.43**	**57.58**	**55.98**
Δ DRA$-$MoRA†	+1.52	+0.90	+0.46	+2.23	+1.70	+1.40
(b) Both-missing scenario; unweighted loss						
MoRA†	54.61	52.15	50.92	56.36	54.99	53.44
DRA	**56.23**	**53.35**	**51.87**	**58.88**	**57.20**	**55.38**
(c) Both-missing scenario; positive-class weight $w = 3$						
MoRA†	57.88	55.36	54.23	58.34	55.87	54.68
DRA	**60.02**	**57.57**	**56.23**	**60.06**	**57.98**	**56.39**

Comparison to DPL. We reimplement DPL's public prototype head and loss (Lu et al., 2026) under the same observed-data protocol. At both-missing 70%, DPL/DRA obtain $55.32/53.35$ fixed F1, $55.42/57.20$ validated F1 and $56.94/59.34$ AP. Across all three scenarios, DPL has higher fixed F1; DRA has higher mean validated F1 and AP (Table 5). DPL trains 705,024 parameters, including final normalization/projections; DRA trains 336,919. The methods retain their respective training recipes (Appendix A.1).

5.3 Decision-rule and recipe controls

Same checkpoints, two rules. DRA's advantage persists under a common validated rule: its paired margin over MoRA† grows from $+0.96$ to $+1.77$, positive in 9 of 9 shared seed pairs (Table 2). Validated per-label thresholds raise F1-Macro by $+2.33$ for MoRA† and $+3.14$ for DRA: Decide is a decision layer usable with any model. Paired margins are computed on shared seed pairs.

Ranking and recipe controls. Across nine MM-IMDb settings, DRA improves AP from 58.02 to 60.23. Giving both methods the same four recipe candidates and checkpoint-selection options yields 58.02 versus 55.82 validated F1 at both-missing 70% (three seeds; Appendix A.2).

Diagnostics (Figure 3a–c). On movie-disjoint MM-IMDb splits, fitted thresholds follow calibrated $\hat{F}_s^\star/2$; the gains concentrate on missing rows. Class weighting shifts logits upward, illustrating why a fixed cutoff can obscure ranking improvements.

5.4 Evidence value and allocation (Read)

State dependence (R1). Three tail windows raise AP by $+1.42$ on image-missing rows and $+0.33$ on complete rows read counterfactually (η=70%); the interaction is $+1.27$ at 90% and near zero at 50% ($+0.02$; Figure 3d). These are finite-reader AP gains: each ratio has a separately trained model and different reader exposure, while MCAR leaves Bayes Δ_s unchanged. To test the effect of selective training, we train the reader on both states and evaluate four paths on the same 2340 originally complete movies. Table 3b gives a log-loss interaction of 0.00839 nats per label ($[0.00767, 0.00905]$), the fitted-model analogue of the positive interaction in Proposition 1(a).

Trained-policy controls. Eligible rows have observed text beyond the prefix; DRA encodes 8334 extra windows on 2778 test rows. Separately trained policies share the architecture and recipe (Table 3a): prefix re-reading changes F1 by -0.70 relative to DRA (DRA higher in 6 of 6 seeds), state-blind reading by -0.59 (higher in 5 of 6), and no reading by -0.69. Reading every eligible row changes F1 by $+0.19$ at $1.87\times$ the extra windows. Added to MoRA† under its official recipe, the reader changes F1 by -0.34 over 3 shared seeds.

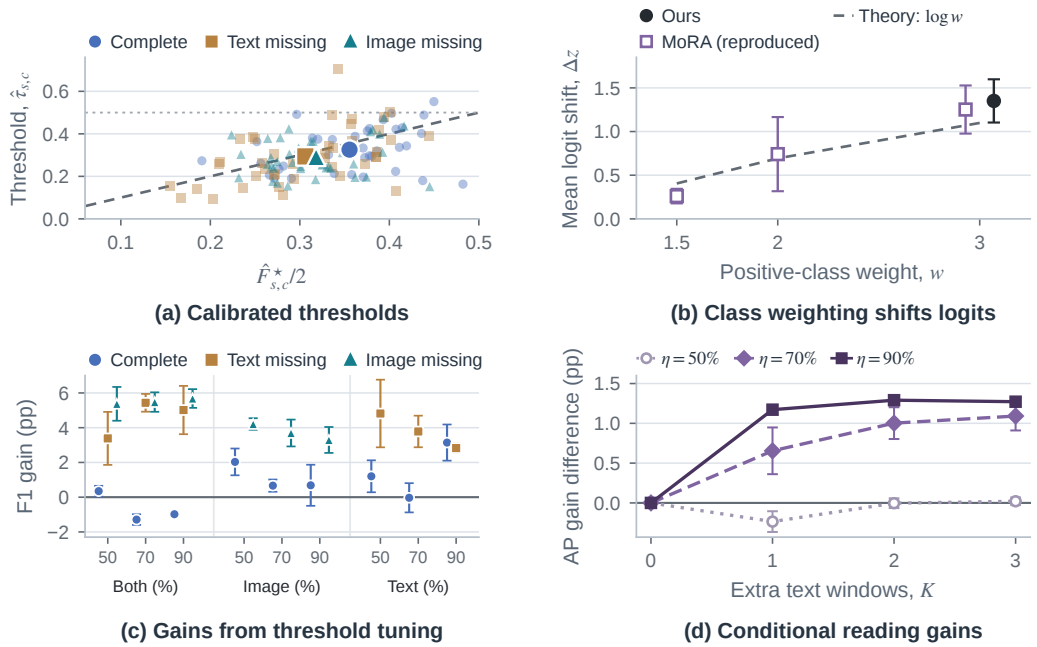


Figure 3: **Decision and reading diagnostics.** (a) Calibrated state-wise thresholds track the predicted $\hat{F}^{\star}_{s}/2$; (b) class weighting shifts logits; (c) threshold fitting helps missing rows most; (d) extra text increases AP more when the image is missing at 70% and 90% missingness. Error bars are seed or setting SD; the protocol is in Appendix A.3.

Fixed-reader allocation (R2). Table 3d tests the ordering in Proposition 1(c): the same reader targets image-missing rows or selects rows without using state, at identical cost. At 8334 encoded windows, targeting improves F1 by $+0.18$ and AP by $+0.33$, with a log-loss advantage of 0.00127 nats per label (movie-cluster 95% interval $[0.00104, 0.00150]$). All three trained seeds favor targeting; two smaller budgets agree. This fixed-reader contrast isolates the allocation effect.

Cost. Merged Adapt adds no measurable overhead: on an RTX 4090 (fp16, batch one), DRA without Read takes 18.64 ms per example versus 19.02 ms for the frozen backbone (Table 8b). At batch 128, Read raises time per example from 0.766 to 0.852 ms ($+11\%$).

5.5 Shared-encoder controls

(A1) The state-private design uses $2.6\times$ the encoder-adapter parameters, changing MM-IMDb F1 by -0.86 and Food101 accuracy by -0.17 (Table 3c). The panel reports encoder-adapter parameters; the shared version has the higher mean on both displayed metrics. Removing the shared or private term, or using parameter-matched LoRA, changes F1 by -0.08, -0.30 and -0.64 (Appendix Table 6). Together with the fixed-reader control, these results support a compact coupled encoder whose mergeability complements the state-guided allocation policy. Fixed-budget rank controls appear in Appendix Table 7.

5.6 Transfer across missing scenarios

Table 4 compares training scenarios (rows) with different test scenarios (columns), without refitting checkpoints or decision rules. DRA exceeds MoRA† in $16/18$ off-diagonal comparisons: $6/6$ on MM-IMDb, $6/6$ on Food101 and $4/6$ on Hateful Memes. MM-IMDb uses macro-AP, measuring ranking transfer; the complete score grid and pure-state tests appear in Appendix Table 9.

Table 3: **Component controls** (both missing, $\eta = 70\%$; seed means). (a) Validated MM-IMDb F1 at extra-window budgets relative to DRA; all-eligible rows use $1.87\times$ budget. Residual uses the learned zero-initialized reader; raw mean averages tail features from identity initialization. (b) Two reader training policies on the same 2340 originally complete movies ($K = 3$, three cross-fitted checkpoints): reductions in unweighted log loss, 10^{-3} nats per label per plot row. Δ is hidden minus present, with movie-cluster 95% CI. (c) Shared versus state-private encoder-adapter parameters, validated MM-IMDb F1 and Food101 accuracy. (d) Targeted minus state-blind allocation at matched cost: F1/AP gains in percentage points; loss advantage is state-blind minus targeted in 10^{-3} nats per label. Bold marks the best displayed value within comparable rows.

(a) Reading policy

Policy	Windows	MM F1	n
MoRA†	0.00×	54.99	3
MoRA† + Read	1.00×	54.28	6
No reading	0.00×	**56.49**	6
Re-read prefix	1.00×	56.48	6
State-blind	1.00×	56.60	6
DRA	1.00×	**57.18**	6
All eligible (residual)	1.87×	**57.37**	6
All eligible (raw mean)	1.87×	57.06	6

(b) Same-movie tail value

Trained on	Present	Hidden	Δ [95% CI]
Image-missing	-2.45	$+7.65$	$+10.11$ $[9.25, 10.91]$
Both states	$+0.60$	$+8.99$	$+8.39$ $[7.67, 9.05]$

(c) One shared vs. state-private encoder

Encoder	Adapter params.	MM F1	Food acc.
Shared encoder[6]	**51.2K**	**57.18**	**84.54**
State-private[3]	133.1K	56.35	84.46

(d) Targeted vs. state-blind allocation

Windows	F1 gain	AP gain	Loss adv.
1,827	$+0.03$	$+0.09$	$+0.28$
3,990	$+0.11$	$+0.18$	$+0.64$
8,334	$+\mathbf{0.18}$	$+\mathbf{0.33}$	$+\mathbf{1.27}$

Table 4: **Cross-scenario transfer:** DRA $-$ MoRA† (percentage points; three-seed means). Rows train, columns test; T/I/B = text/image/both missing. Green favors DRA, red favors MoRA†. Gray diagonal is in-domain (Tables 1, 2); full scores: Appendix Table 9. No target refitting.

MM-IMDb (AP)

	T	I	B
T	–	$+0.39$	$+1.50$
I	$+6.80$	–	$+4.63$
B	$+3.51$	$+1.29$	–

Food101 (acc.)

	T	I	B
T	–	$+3.86$	$+2.64$
I	$+5.59$	–	$+2.75$
B	$+1.88$	$+0.42$	–

Hateful Memes (AUROC)

	T	I	B
T	–	$+0.65$	$+1.28$
I	$+1.27$	–	-0.54
B	$+1.78$	-0.55	–

5.7 Limitations

Read's allocation evidence is limited to MM-IMDb; Food101 and Hateful Memes evaluate the shared encoder under accuracy and AUROC. The theory concerns Bayes risks and state-level policies under MCAR; learned-reader comparisons establish empirical signs and modest allocation gains, not Bayes optimality. Additional controls and implementation details are in Appendix A.3.

6 Conclusion

DRA uses missingness to allocate computation to observed evidence. The analysis characterizes loss-dependent reading priorities and the boundary where state-blind allocation is optimal. On MM-IMDb, same-movie controls with symmetric training support a log-loss substitute pattern, and fixed-reader comparisons show a modest advantage at identical budgets. A state-gated residual reader implements this principle using shared adaptation and a common evaluation interface.

AI USE STATEMENT

We used Claude Code and Codex to aid and polish the language and presentation, consistent with the submission declaration. The authors designed and ran the experiments, verified the derivations and statistical calculations, and take responsibility for the final methods, code, results, and text.

ETHICS STATEMENT

Experiments use three public benchmarks and collect no new human-subject data. Hateful Memes contains offensive material; we use it for evaluation without reproducing examples. We study controlled missingness and make no deployment claim for content moderation or other consequential settings.

REPRODUCIBILITY STATEMENT

Section 3, Section 4, Table 10 and Appendix B specify the assumptions, model, training configuration and evaluation protocol. We verified the derivations and recomputed paired intervals from the saved predictions and movie clusters. Tables are generated from saved per-seed results; code, configurations and saved predictions will be released.

REFERENCES

John Arevalo, Thamar Solorio, Manuel Montes-y-Gómez, and Fabio A. González. Gated multimodal units for information fusion. In *International Conference on Learning Representations Workshops (ICLRW)*, 2017.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (COLT)*, pp. 92–100, 1998. doi: 10.1145/279943.279962.

Tilman Börgers, Angel Hernando-Veciana, and Daniel Krähmer. When are signals complements or substitutes? *Journal of Economic Theory*, 148(1):165–195, 2013. doi: 10.1016/j.jet.2012.12.012.

Taixi Chen and Nancy Guo. Test-time logit prompting for source-free missing modality adaptation. *arXiv preprint arXiv:2609.02039*, 2026.

A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15, 1979. doi: 10.1111/j.2517-6161.1979.tb01052.x.

Charles Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 973–978, 2001.

Yubo Gao, Haotian Wu, Xiaoyu Xu, Yibo Yan, Hong Chen, Ruoshui Peng, Fei Pan, Puay Siew Tan, Zhuoran Gao, Yonghua Hei, Jie Zhang, and Xuming Hu. Should missing modalities always be necessary to repair for multi-modal sentiment analysis? *arXiv preprint arXiv:2607.17262*, 2026.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Lianyu Hu, Tongkai Shi, Wei Feng, Fanhua Shang, and Liang Wan. Deep correlated prompting for visual recognition with missing modalities. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Jaehyuk Jang, Yooseung Wang, and Changick Kim. Towards robust multimodal prompting with missing modalities. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

Jaromír Janisch, Tomáš Pevný, and Viliam Lisý. Classification with costly features using deep reinforcement learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2023.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Oluwasanmi Koyejo, Nagarajan Natarajan, Pradeep Ravikumar, and Inderjit S. Dhillon. Consistent binary classification with generalized performance metrics. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

Oluwasanmi O. Koyejo, Nagarajan Natarajan, Pradeep K. Ravikumar, and Inderjit S. Dhillon. Consistent multilabel classification. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2015.

Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2023.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Russ R. Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Zachary C. Lipton, Charles Elkan, and Balakrishnan Narayanaswamy. Optimal thresholding of classifiers to maximize F1 measure. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2014.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning (ICML)*, 2018.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *International Conference on Machine Learning (ICML)*, 2024.

Jueqing Lu, Yuanyuan Qi, Xiaohao Yang, Shuaicheng Niu, Fucai Ke, Shujie Zhou, Wei Tan, Jionghao Lin, Wray Buntine, Hamid Rezatofighi, and Lan Du. DPL: Decoupled prototype learning for enhancing robustness of vision-language transformers to missing modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39219–39229, 2026.

William J. McGill. Multivariate information transmission. *Psychometrika*, 19(2):97–116, 1954.

Harikrishna Narasimhan, Rohit Vaish, and Shivani Agarwal. On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

Natalia Neverova, Christian Wolf, Graham W. Taylor, and Florian Nebout. ModDrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(8):1692–1706, 2016.

Raghavendra Pappagari, Piotr Żelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. Hierarchical transformers for long document classification. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

Ignazio Pillai, Giorgio Fumera, and Fabio Roli. Threshold optimisation for multi-label classifiers. *Pattern Recognition*, 46(7):2055–2065, 2013.

John C. Platt. Probabilities for SV machines. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans (eds.), *Advances in Large-Margin Classifiers*, pp. 61–74. MIT Press, 2000.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.

Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A. Smith. The right tool for the job: Matching model and instance complexities. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune BERT for text classification? In *China National Conference on Chinese Computational Linguistics (CCL)*, 2019.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso. Recipe recognition with large multimodal food dataset. In *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2015.

Shicai Wei, Chunbo Luo, and Yang Luo. MMANet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2023.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *Transactions on Machine Learning Research (TMLR)*, 2026.

Zhixiang Xu, Kilian Q. Weinberger, and Olivier Chapelle. The greedy miser: Learning under test-time budgets. In *International Conference on Machine Learning (ICML)*, 2012.

Adams Wei Yu, Hongrae Lee, and Quoc Le. Learning to skim text. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

Bianca Zadrozny, John Langford, and Naoki Abe. Cost-sensitive learning by cost-proportionate example weighting. In *IEEE International Conference on Data Mining (ICDM)*, pp. 435–442, 2003.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-CLIP: Unlocking the long-text capability of CLIP. In *European Conference on Computer Vision (ECCV)*, 2024a.

Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. DePT: Decoupled prompt tuning. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2024b.

Zhihui Zhang, Luanyuan Dai, Qika Lin, Yunfeng Diao, Guangyin Jin, Yufei Guo, Jing Zhang, and Xiaoshuai Hao. Synergistic prompting for robust visual recognition with missing modalities. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Table 5: **DPL comparison under a common observed-data protocol.** MM-IMDb, $\eta = 70\%$, three seeds; means with seed SD in parentheses. DPL: public prototype/loss implementation and trainable final normalization/projections. DRA: unweighted main-table checkpoints. Both use the same validation-threshold procedure. Bold: higher mean within each scenario and metric. Training recipes and parameter budgets differ (Appendix A.1).

Scenario	Method	Fixed F1	Validated F1	Macro-AP
Both missing	DPL	**55.32** (0.28)	55.42 (0.19)	56.94 (0.12)
	DRA	53.35 (0.34)	**57.20** (0.10)	**59.34** (0.04)
Image missing	DPL	**57.59** (0.40)	57.92 (0.27)	60.46 (0.15)
	DRA	56.81 (0.10)	**59.81** (0.63)	**62.82** (0.62)
Text missing	DPL	**55.03** (0.45)	55.53 (0.42)	57.00 (0.14)
	DRA	52.72 (0.39)	**55.74** (0.80)	**58.03** (0.71)

Shu Zhao, Nilesh Ahuja, Tan Yu, Tianyi Shen, and Vijaykrishnan Narayanan. MoRA: Missing modality low-rank adaptation for visual recognition. In *International Conference on Learning Representations (ICLR)*, 2026.

Xinkui Zhao, Jinsong Shu, Yangyang Wu, Guanjie Cheng, Zihe Liu, Naibo Wang, Shuiguang Deng, Zhongle Xie, and Jianwei Yin. A robust incomplete multimodal low-rank adaptation approach for emotion recognition. *arXiv preprint arXiv:2507.11202*, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.

A Additional experimental results

A.1 Direct-neighbor diagnostic

We reimplement DPL's prototype head, ArcFace terms and prototype contrastive loss from its CVPR 2026 public code. Both-missing validation fixed F1 selects $\lambda = 0.5$ from $\{0.2, 0.5\}$, which is then fixed across scenarios. We train for 20 epochs with AdamW (0.01 learning rate, 0.02 weight decay, 10% warm-up, linear decay), selecting the best validation fixed F1 at five checkpoints per epoch. The public freeze predicate retains the final vision/text LayerNorm and both CLIP projections ($705{,}024$ trainable parameters including prototypes); cached and direct features agree numerically.

Table 5 compares all three scenarios at $\eta = 70\%$. Both methods share CLIP weights, preprocessing, masks, rows, labels and threshold fitting, and never receive hidden modalities; DRA uses its unchanged unweighted checkpoints. DPL retains its public loss and recipe, with cached features used for the paired evaluation. DPL has higher fixed F1; DRA has higher mean validated F1 and AP in every scenario.

A.2 MoRA reproduction and recipe controls

The official-code MoRA rerun retains MoRA's official adapter, optimiser and loss recipe. It shares DRA's released split masks, seeds, evaluation rules and checkpoint-selection protocol, so the paired comparisons do not change the data or selection criteria; numerical hyperparameters remain method-specific and are listed in Table 10. Its three-scenario means and paired differences are reported in Table 2 and the component controls. The 27 settings reuse test examples, and recipes and selection criteria are in Section 5.1 and Table 10.

Loss selection for the decision control. For Table 2, both models select among unweighted BCE, positive weights 1.5, 2, 3 and an asymmetric loss on the both-missing validation split; both select weight 3. Thresholds are fitted separately by model, ratio, scenario and label using Section 4.3's common procedure.

Table 6: **Encoder component controls.** Both missing, $\eta = 70\%$; MM: validated MM-IMDb F1; Food: accuracy; HM: AUROC. Superscripts give training-seed counts. Terms are removed one at a time under the DRA recipe; MoRA† retains its official recipe. Bold marks the highest displayed mean within each metric, without implying a significant difference.

Encoder	MM	Food	HM
DRA^6	$\mathbf{57.18}$	84.54	69.71
w/o shared^6	57.10	84.52	69.78
w/o private^6	56.89	84.49	69.10
LoRA (matched)^3	56.56	$\mathbf{84.58}$	69.52
Frozen adapters^6	56.38	84.39	68.69
MoRA$^{\dagger\,3}$	54.99	83.52	$\mathbf{70.48}$

Table 7: **Fixed-budget rank allocation.** MM-IMDb both-missing 70%, validated F1 means over three seeds; each allocation uses $30{,}720$ encoder parameters. †: allocation chosen on validation data. Bold: best displayed mean per column.

Shared / private rank	Official selector	Cross-fitted selector
4 / 2†	57.17	57.92
8 / 1	$\mathbf{57.34}$	$\mathbf{58.11}$
0 / 3	57.12	57.97

Common recipe search. The four candidates are each method's base recipe, classifier learning rate 0.003, EMA decay 0.99, or adapter learning rate divided by three. Both also choose official-metric versus movie-grouped cross-fitted checkpoints using three-seed validation means. MoRA selects EMA, DRA the classifier learning rate, and both cross-fitted checkpoints; validated F1 is 58.02 versus 55.82. Search categories and rules match while base configurations remain method-specific.

A.3 Checkpoint selection and inference cost

Checkpoint selection. Table 8a repeats the reading ablation with movie-grouped cross-fitted checkpoint selection. DRA remains above no reading, prefix re-reading and state-blind reading; reading every row is slightly higher at $1.87\times$ the extra windows.

Diagnostic protocol for Figure 3. For Figure 3, (a) fits state/label Platt maps on one movie partition and estimates thresholds on the other with swapped roles; points average three seeds. In (b), markers are mean (sample SD) logit shifts; DRA uses $w = 3$, while $w = 1.5, 2$ are MoRA† controls. In (c), thresholds are cross-evaluated on movie-disjoint folds; (d) contrasts image-missing with counterfactual complete-row AP gains. Panels (a–c) use official checkpoints, (d) cross-fitted checkpoints; error bars are seed SD. Panel (d) uses distinct checkpoints at each ratio. The image-missing state has probability $\eta/2$ in the both-missing protocol; the training gate reads only its valid-tail rows. Checkpoint selection and evaluated state cohorts also vary with η. This diagnostic does not isolate training exposure as the cause of the changing AP interaction.

Initialization and threshold endpoints. For the additive encoder, each shared factor is sampled independently from $\mathcal{N}(0, 0.01^2)$, each private left factor from the same distribution, and each private right factor is zero; the residual scale is α/r_m. The linear head uses the upstream normal initialization with standard deviation 0.02. The image-missing, re-read and state-blind reader projections are zero-initialised, so they are exact no-ops at step zero; the raw-mean reference uses identity initialization. These settings are recorded in the released configuration schema. Decide evaluates the distinct validation-score cuts in descending order, including the empty and all-positive endpoints; ties use the larger threshold, and the all-positive endpoint is represented by the strict floating-point predecessor of the minimum validation score. This order-defined construction makes the monotonic-score argument exact.

Table 8: **Reading robustness and inference cost.** (a) Cross-fitted checkpoint selection for the controls of Table 3a; mean validated F1, n seeds, relative window budgets; bold: best at each budget. (b) Forward-pass milliseconds per example on an RTX 4090, fp16, ViT-B/16, with MM-IMDb both-missing $\eta = 70\%$ inputs. Batch 1 averages per-row medians over three rounds on 600 rows; batch 128 uses the median amortised time over three rounds of 16 batches. Loading and data transfer are excluded; bold: lowest displayed time per column.

(a) Checkpoint selection

Policy	Windows	MM F1	n
MoRA†	0.00×	55.44	3
MoRA† + Read	1.00×	55.65	6
No reading	0.00×	**57.55**	6
Re-read prefix	1.00×	57.26	6
State-blind	1.00×	57.84	6
DRA	1.00×	**58.06**	6
All eligible (residual)	1.87×	**58.28**	6
All eligible (raw mean)	1.87×	58.26	6

(b) Inference cost

Model	Batch 1	Batch 128
Frozen CLIP + head	19.02	**0.763**
MoRA	26.44	0.768
MoRA, merged	18.82	**0.763**
Adapt	21.13	0.785
Adapt, merged	**18.64**	0.766
Adapt + Read	25.09	0.884
Adapt + Read, merged	20.53	0.852

Matched reading with symmetric training exposure. Table 3b evaluates four paths on identical movies. The second reader is the all-row residual control from Table 3a, trained on both observed-text states; its positive gain difference addresses selective exposure. Rows without a valid tail contribute zero to both gains, so the all-movie average preserves the eligible-cohort interaction sign. Intervals condition on the three trained seeds and movie clusters.

Matched allocation with a symmetric-trained reader. Table 3d fixes three cross-fitted all-row residual checkpoints and $K = 3$. In buckets of valid-tail count and capped tokens, each policy selects 25%, 50% or 100% of image-missing rows after rounding down; targeting restricts candidates to that state, while state-blind uses all observed-text rows. The 20 fixed label-free hash seeds give matched encoded-window budgets 1,827, 3,990, 8,334 (valid windows 903, 2,024, 4,281). Thresholds are fit separately on validation data; loss intervals use 2,000 movie-cluster bootstrap draws over paired allocation/seed differences and condition on the three trained models. Allocation repeats are not training seeds.

Table 9 reports the three-seed means underlying the non-diagonal margins in Table 4. The pure-state columns evaluate each trained model with every test row forced into one observed state; no target-scenario checkpoints or thresholds are selected.

B BENCHMARKS, MISSING-STATE PROTOCOL AND METRICS

Movie grouping. The official MM-IMDb loader scores every plot summary separately (Zhao et al., 2026). Diagnostic folds keep all plots from a movie together. Food101 pairs images with web-page text (Wang et al., 2015); Hateful Memes pairs meme images with overlaid text (Kiela et al., 2020).

Training information. All splits use the released missing-state tables. Hidden modalities supply neither model inputs nor teacher or reconstruction targets.

Aggregation. MM-IMDb uses the official TorchMetrics macro-F1 over plot rows and 23 labels. Scenario averages give equal weight to image-missing, text-missing and both-missing results at the same ratio; these settings reuse test examples.

Table 9: **Full cross-scenario and pure-state scores.** Each pair of rows is trained at $\eta = 70\%$ under the named scenario. The left block tests all three scenario tables: off-diagonal cells are unseen training-to-test transfers. The right block forces the test split into one state. Scores are three-seed means with no target-scenario refitting; MM-IMDb uses macro-AP. Bold compares DRA with MoRA† at the same train/test setting.

		Tested on scenario			Tested on pure state		
Trained on	Method	Text miss.	Image miss.	Both miss.	Complete	Image only	Text only
MM-IMDb (macro-AP)							
Text missing	MoRA†	56.40	51.51	52.61	64.92	52.55	**47.57**
	DRA	**58.03**	**51.90**	**54.11**	**65.60**	**54.50**	47.50
Image missing	MoRA†	41.97	60.53	50.59	62.32	32.16	60.09
	DRA	**48.77**	**62.82**	**55.21**	**65.31**	**40.90**	**61.18**
Both missing	MoRA†	53.95	60.65	56.78	66.10	48.07	58.51
	DRA	**57.46**	**61.93**	**59.34**	**67.76**	**52.74**	**59.29**
UPMC-Food101 (accuracy)							
Text missing	MoRA†	80.41	79.36	80.14	93.24	75.25	73.26
	DRA	**81.76**	**83.21**	**82.78**	**93.83**	**76.90**	**78.51**
Image missing	MoRA†	62.35	88.11	75.39	92.62	49.48	**86.09**
	DRA	**67.95**	**88.31**	**78.14**	**93.16**	**57.11**	**86.09**
Both missing	MoRA†	78.86	87.68	83.52	93.19	73.06	85.28
	DRA	**80.74**	**88.10**	**84.62**	**94.13**	**75.35**	**85.39**
Hateful Memes (AUROC)							
Text missing	MoRA†	68.99	64.44	68.72	71.76	68.24	63.76
	DRA	**71.46**	**65.09**	**70.00**	**73.19**	**70.45**	**64.70**
Image missing	MoRA†	64.31	**69.66**	**67.84**	72.13	60.80	**69.16**
	DRA	**65.58**	69.14	67.30	**73.01**	**64.51**	68.25
Both missing	MoRA†	67.05	**68.21**	**70.48**	71.75	64.82	**68.07**
	DRA	**68.84**	67.65	69.71	**71.85**	**67.65**	67.54

Table 10: **Configuration of DRA.** All datasets use AdamW (β=$(0.9, 0.98)$) with a cosine schedule and 10% warm-up, mixed precision and the checkpoint selected by the official validation metric. Recipes were selected on the validation split and are shared by all settings of a dataset.

	MM-IMDb	UPMC-Food101	Hateful Memes
Adapted projections (last two layers)	query, key	query, value	key, value
Rank (shared / per modality)	4 / 4	4 / 4	8 / 8
Scaling α / adapter dropout	8 / 0.1	8 / 0.1	8 / 0.1
Adapter / classifier learning rate	3×10^{-4} / 3×10^{-3}	3×10^{-4} / 3×10^{-3}	10^{-3} / 3×10^{-3}
Weight decay	0.02	0.02	0.1
Batch size / epochs	256 / 20	256 / 20	16 / 2
Tail windows K (image missing)	3	–	–

C Proofs and remarks

C.1 Value and allocation of unread evidence

Proof of Proposition 1. All quantities in this proof use the conditional law of rows with a valid tail. Missingness remains independent of (X, Y) under this restriction.

(a) Decision-dependent substitutes. Expanding the risk reductions gives

$$\begin{aligned}\Delta_{\mathrm{T}} - \Delta_{\mathrm{F}} &= H_\ell(Y \mid P) - H_\ell(Y \mid P, U) - H_\ell(Y \mid P, V) + H_\ell(Y \mid P, U, V) \\ &= I_\ell(Y; V \mid P) - I_\ell(Y; V \mid P, U).\end{aligned}$$

Thus either view reduces the other's marginal value by the same amount. A nonnegative difference defines substitutes for the specified prior, loss and conditioning information; a nonpositive difference

Table 11: Benchmark splits. For MM-IMDb the first row counts movies and the second counts plot rows; each plot is scored as a separate row that shares its movie's poster, labels and missing state.

Dataset	Task	Labels	Train	Validation	Test	Metric
MM-IMDb (movies)	multi-label	23	15,552	2,608	7,799	macro-F1
plot rows			32,278	5,411	16,120	
UPMC-Food101	single-label	101	61,127	6,845	22,716	top-1 accuracy
Hateful Memes	binary	2	8,500	500	1,000	AUROC

defines complements, with equality at the boundary (Börgers et al., 2013). This characterization does not assert substitutability for every decision problem. Finite risks exclude undefined differences.

(b) Conditional independence and a loss-dependent reversal. For log loss, the Bayes risk is conditional entropy. Expanding $I(U; Y, V \mid P)$ in two orders yields

$$I(U; V \mid P) + I(U; Y \mid V, P) = I(U; Y \mid P) + I(U; V \mid Y, P).$$

Hence $\Delta_{\mathrm{T}} - \Delta_{\mathrm{F}} = I(U; V \mid P) - I(U; V \mid P, Y)$, the conditional co-information (McGill, 1954), expressing net redundancy versus synergy (Williams & Beer, 2010; Liang et al., 2023). Under $U \perp V \mid (P, Y)$ the second term vanishes, so the views are substitutes for log loss. This uses the conditional-independence assumption familiar from co-training (Blum & Mitchell, 1998), without requiring that each view determine the label.

The sign can reverse without conditional independence. With constant P, if $U = V = Y$ is a fair bit, then $\Delta_{\mathrm{T}} = \log 2$ and $\Delta_{\mathrm{F}} = 0$. If U, V are independent fair bits and $Y = U \oplus V$, then $\Delta_{\mathrm{T}} = 0$ and $\Delta_{\mathrm{F}} = \log 2$.

Even with conditional independence, the loss matters. Let P be constant, $\Pr(Y = 1) = 0.8$, and U, V be independent given Y, each equal to Y with probability 0.7. With no view or either single view, the Bayes 0–1 decision is always 1, since even $\Pr(Y = 1 \mid U = 0) = 0.24/(0.24 + 0.14) > 1/2$. The risk is 0.2, so $\Delta_{\mathrm{T}} = 0$. With both views, the decision changes to 0 exactly on $\{U = V = 0\}$: the joint masses for $Y = 0$ and $Y = 1$ there are 0.098 and 0.072. This lowers risk to $0.2 - 0.098 + 0.072 = 0.174$, giving $\Delta_{\mathrm{F}} = 0.026$. The views are therefore strict complements under 0–1 loss. For log loss they are strict substitutes: conditional independence gives a difference of $I(U; V) > 0$, since $\Pr(U = V = 1) = 0.41 \neq 0.62^2 = \Pr(U = 1) \Pr(V = 1)$.

(c) Optimal state allocation and the state-blind boundary. Let R indicate reading, drawn with probability ρ_S independently of (X, Y) given S, and observed by the predictor. The state-s Bayes risk after allocation is

$$(1 - \rho_s) H_\ell(Y \mid h_s(X)) + \rho_s H_\ell(Y \mid h_s(X), U).$$

Thus reading reduces overall risk by $\sum_s p_s \rho_s \Delta_s$ at cost $\sum_s p_s \rho_s c_s$. Each $\Delta_s \geq 0$, because a predictor can ignore U. This is a fractional knapsack: write $b_s = p_s c_s \rho_s$ and $r_s = \Delta_s / c_s$; the objective is $\sum_s b_s r_s$ with $0 \leq b_s \leq p_s c_s$ and $\sum_s b_s \leq B$. If $r_i > r_j$, shifting cost $\epsilon > 0$ from j to an unsaturated i improves it by $\epsilon(r_i - r_j)$. Reading in decreasing r_s is therefore optimal, with at most one partially read state, until all useful reads or the budget are exhausted.

For the two states, suppose $r_{\mathrm{T}} > r_{\mathrm{F}}$ and a feasible state-blind rule uses common probability ρ. Since $B < \sum_s p_s c_s$, $\rho < 1$. If $\rho > 0$, then $p_{\mathrm{T}}, p_{\mathrm{F}} > 0$ allow a cost-preserving exchange from F to T that strictly improves the objective; if $\rho = 0$, a small read in T improves it because $B > 0$ and $r_{\mathrm{T}} > 0$. The opposite ordering is symmetric. Conversely, when the ratios agree, the budget-matched common probability $B / \sum_s p_s c_s$ is optimal. With equal costs this boundary is $\Delta_{\mathrm{T}} = \Delta_{\mathrm{F}}$. For mean binary log loss over labels, use their mean gain as Δ_s. The result optimizes state-level randomization, with fixed gains and costs, rather than policies that select examples within a state. □

Remark: finite readers and missing ratios. Under MCAR, the population risks defining Δ_s are independent of the state probabilities and hence of the missing ratio η. For predictors trained at η, let $\varepsilon_s^P(\eta)$ and $\varepsilon_s^{P,U}(\eta)$ denote their excess population risks over the corresponding Bayes risks. Their expected reading gain is

$$\widetilde{\Delta}_s(\eta) = \Delta_s + \varepsilon_s^P(\eta) - \varepsilon_s^{P,U}(\eta).$$

Training exposure and approximation error can therefore change a finite reader's gain while Δ_s stays fixed; empirical estimates also incur sampling error. Figure 3d measures AP changes in distinct

state cohorts using separately trained checkpoints at each η, whereas Table 3b estimates a log-loss interaction on the same movies. Its positive interaction under symmetric training supports a substitute relation for these fitted predictors, without identifying the Bayes interaction or establishing conditional independence.

C.2 Garbling and the F1 operating point

Let $\mu(O) = \Pr(Y = 1 \mid O)$ and $\pi = \Pr(Y = 1) > 0$. For a randomized decision rule $\delta : \mathcal{O} \to [0, 1]$ that predicts positive independently of Y given O,

$$F(\delta) = \frac{2\mathbb{E}[\delta(O)Y]}{\mathbb{E}[\delta(O)] + \pi} = \frac{2\mathbb{E}[\delta(O)\mu(O)]}{\mathbb{E}[\delta(O)] + \pi}. \tag{6}$$

Lemma 1 (F1 cannot improve under garbling)**.** *If O' is obtained from O by a label-independent Markov kernel, then $F^\star(O') \le F^\star(O)$, where the optima can be taken over deterministic rules. In Eq. (1), increasing the missing ratio within any of the three fixed missing-type protocols cannot increase $F^\star$.*

Proof. We first justify deterministic attainment for this nondecomposable metric. Following the optimal-cut characterization of Lipton et al. (2014), let $G(f) = \mathbb{E}[(2\mu - f)_+] - f\pi$. It is continuous and strictly decreasing on $[0, 1]$, with $G(0) > 0$ and $G(1) \le 0$, so it has a unique zero $f^\star$. For every randomized rule,

$$2\mathbb{E}[\delta\mu] - f^\star(\mathbb{E}[\delta] + \pi) \le G(f^\star) = 0,$$

and $\delta^\star = \mathbf{1}\{\mu \ge f^\star/2\}$ attains equality. Thus $f^\star = F^\star(O)$ is also the optimum among randomized rules. A rule $\delta'(O')$ is simulated from O by $\delta(O) = \mathbb{E}[\delta'(O') \mid O]$. Since $Y \perp O' \mid O$, this preserves both expectations in Eq. (6) and hence F1, proving the inequality.

For missing ratios $\eta < \eta'$, independently erase the specified modality from each complete row with probability $(\eta' - \eta)/(1 - \eta)$ in either single-missing-type protocol. In the both-missing protocol, erase the image or the text with probability $(\eta' - \eta)/(2(1 - \eta))$ each, mutually exclusively. These garblings transform the state probabilities at η into those at η'. □

By Lipton et al. (2014), the canonical cutoff is $F^\star/2$. Since P and V are each garblings of (P, V) and MCAR preserves their label law, $F^\star_{\mathrm{T}}, F^\star_{\mathrm{I}} \le F^\star_{\mathrm{F}}$, and the canonical state cutoffs inherit this weak ordering. For pooled F1, O includes S and the optimal rule compares its posterior to the same $F^\star(O)/2$ in every state. State-wise validation optima $\hat{F}^\star_s$ serve only as diagnostics; deployed Decide fits pooled per-label thresholds.

Remark: threshold nonuniqueness and regret. Let $q = \mu(O)$, $t^\star = F^\star/2$ and $A^\star = \{q \ge t^\star\}$. For any decision set A,

$$F^\star - F(A) = \frac{2\mathbb{E}[|q - t^\star|\mathbf{1}_{A \triangle A^\star}]}{\Pr(A) + \pi}.$$

Multiplying by the denominator and substituting $2\mathbb{E}[q\mathbf{1}_{A^\star}] = F^\star(\Pr(A^\star) + \pi)$ gives $2\mathbb{E}[(q - t^\star)(\mathbf{1}_{A^\star} - \mathbf{1}_A)]$, the displayed numerator. A lower cutoff alone need not increase fixed-threshold regret. For example, let q take values $(0.01, 0.41, 0.85)$ with probabilities $(0.2, 0.4, 0.4)$. Optimal F1 is 0.77182, while the fixed-0.5 rule gives 0.75055. Garbling to a constant yields posterior 0.506 and optimal F1 0.67198, now attained by the default rule: the optimum and regret both decrease. Either decision at $q = t^\star$ is optimal, and distinct cutoffs can induce identical decisions.

C.3 Classical facts used

Positive-class weighting $w > 0$ gives the unrestricted weighted-BCE optimum $\sigma(z_w) = w\mu/(w\mu + 1 - \mu)$, hence $z_w = \operatorname{logit}\mu + \log w$ for $0 < \mu < 1$ (Elkan, 2001; Zadrozny et al., 2003). Strictly increasing score transforms preserve validation-score order, ties and order-defined endpoint cuts; applying the same transform to test scores therefore preserves the selected predictions under fixed tie-breaking (Section 4.3). Finally, for $Z = g(X_m)$, joint sufficiency $X_m \perp (Y, X_p) \mid Z$ implies $Y \perp X_m \mid (Z, X_p)$ and $Y \perp X_m \mid Z$ by weak union and decomposition (Dawid, 1979). This classical condition is not assumed for the learned encoder; its sharing design is evaluated by (A1).