BALANCING MULTIMODAL LEARNING VIA SELECTIVE KNOWLEDGE TRANSFER WITH PATH DISCREPANCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Modality imbalance remains a fundamental challenge in multimodal learning, as heterogeneous modalities exhibit disparate learning dynamics and contribute unequally to joint optimization. Existing methods primarily improve learning balance through optimization regulation, while knowledge transfer complements these efforts by strengthening under-optimized modalities with multimodal supervision, typically through representation alignment. However, alignment-based transfer uniformly reduces cross-modal discrepancy without distinguishing knowledge deficiency caused by insufficient modality learning from modality-conditioned variation arising through cross-modal interaction, thereby conflating information that should be transferred with information that should remain distinctive. We propose Path-Discrepancy Guided Selective Transfer (PDST), which exploits transfer-path consistency to determine how different cross-modal knowledge should be utilized for balanced multimodal learning. By comparing direct transfer with its modality-conditioned counterpart under identical source and target spaces, PDST identifies route-stable information as path-invariant knowledge and interaction-induced variation as path-dependent knowledge. The resulting discrepancy-guided mechanism selectively transfers path-invariant knowledge to compensate for insufficient modality learning while preserving path-dependent knowledge to maintain complementary cross-modal information. Extensive experiments on multiple multimodal datasets demonstrate consistent improvements over representative balanced multimodal learning methods, establishing selective knowledge transfer guided by path discrepancy as an effective approach to mitigating modality imbalance.

1 INTRODUCTION

Multimodal learning integrates complementary information across heterogeneous modalities (Ngiam et al., 2011). Joint training, however, does not guarantee a performance advantage over unimodal models (Wang et al., 2020). Modality competition and preferential learning can cause dominant branches to suppress weaker ones, leaving their representations under-optimized (Huang et al., 2022; Wu et al., 2022). As illustrated in Figure 1(a), such modality imbalance limits the effective contribution of individual modalities even when multimodal fusion improves overall prediction, motivating balanced multimodal learning as an important problem in multimodal representation (Peng et al., 2022; Wei et al., 2024b).

Existing studies regulate modality-specific optimization through gradient modulation (Peng et al., 2022; Li et al., 2023), unimodal learning states and capacity (Wei et al., 2024b), optimization interference (Zhang et al., 2024), or compatibility between unimodal assistance and multimodal objectives (Wei & Hu, 2024). Knowledge transfer complements optimization regulation by supplying supervision beyond each modality's task signal. Multimodal representations are useful sources because they capture both unimodal information and paired information arising from cross-modal interaction (Du et al., 2023). Cross-modal distillation exploits multimodal knowledge to improve target-modality learning (Xue et al., 2023; Huo et al., 2024).

Cross-modal transfer commonly aligns source and target predictions or representations. Modality focusing motivates transferring target-relevant knowledge (Xue et al., 2023), while customized distillation mechanisms account for the modality gap (Huo et al., 2024). Yet endpoint discrepancy does not distinguish knowledge absent from an under-optimized branch from paired information

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.

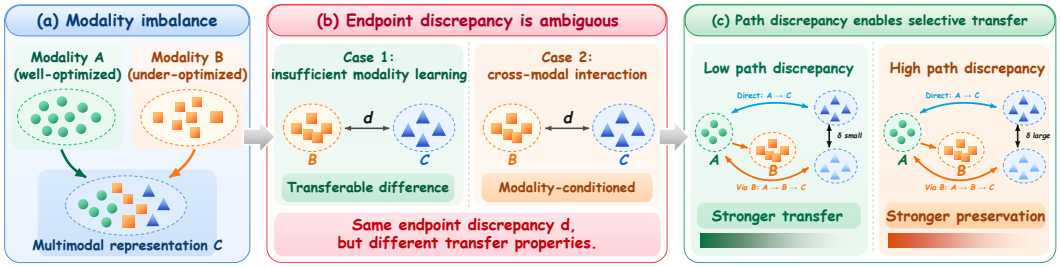


Figure 1: Modality imbalance and selective knowledge transfer. (a) Unequal optimization can leave a modality under-optimized. (b) Endpoint discrepancy conflates missing knowledge with interaction-dependent information. (c) Path discrepancy distinguishes route-consistent supervision from variation induced by modality mediation.

arising through cross-modal interaction (Du et al., 2023). Figure 1(b) illustrates this ambiguity: missing knowledge can provide useful supervision, whereas interaction-dependent variation may be unavailable to an isolated modality. Treating both as transferable obscures which information should be aligned and which should remain distinctive.

We characterize transferability by comparing alternative transfer paths. In Figure 1(c), knowledge from modality A reaches multimodal space C directly or through modality B. Shared endpoints isolate the effect of modality mediation: agreement identifies path-invariant knowledge, while disagreement captures path-dependent variation. We propose *Path-Discrepancy Guided Selective Transfer* (PDST), which constructs both routes in an auxiliary knowledge space and distills their consensus. Our contributions are: (1) a transfer-path formulation separating route-consistent knowledge from modality-conditioned variation; (2) a selective-transfer mechanism using path agreement to construct the teacher and path discrepancy to determine supervision strength; and (3) evaluation on six benchmarks, with ablations and representation analyses linking performance gains to target selection and adaptive weighting. PDST strengthens modality-specific learning while retaining the original inference pathway.

2 Related Work

2.1 Balanced Multimodal Learning

Balanced multimodal learning addresses unequal learning progress and modality dominance through several complementary mechanisms. Gradient modulation adjusts modality-specific updates in OGM and AGM (Peng et al., 2022; Li et al., 2023), while MMPareto reconciles unimodal and multimodal objectives (Wei & Hu, 2024). Representation-level approaches stimulate weaker modalities using class prototypes in PMR (Fan et al., 2023) or sample-specific contribution estimates in SMV (Wei et al., 2024a). Alternating adaptation in MLA (Zhang et al., 2024) and boosting-based reconciliation in ReconBoost (Hua et al., 2024) reduce interference between modality learners; adaptive subnetwork masking offers a complementary intervention within each network (Yang et al., 2025a).

Recent approaches regulate category-level learning difficulty through DLMG/LFM (Yang et al., 2024), sample ordering through BSS (Guan et al., 2025), intervention timing through IPRM (Yang et al., 2025b), and gradient plasticity through IGM (Jiang et al., 2025a). AUG strengthens weaker classifiers to reduce classification-ability disproportion (Jiang et al., 2025b), while MCUT coordinates unimodal assistance (Pan & Yang, 2025) and TCMax promotes multimodal dependence through total correlation (Yu et al., 2026). Across these approaches, balancing primarily concerns optimization or modality capacity. PDST instead selects transferable knowledge by separating route-consistent supervision from modality-conditioned variation.

2.2 Cross-Modal Interaction and Knowledge Transfer

Knowledge distillation uses teacher predictions or representations to supervise a student (Hinton et al., 2015; Gou et al., 2021). Early cross-modal distillation transferred representation-level supervision from labeled RGB images to paired depth and optical-flow inputs (Gupta et al., 2016). Multimodal

teachers additionally encode unimodal and paired information (Du et al., 2023), making the relevance of transferred content to the target modality consequential. Modality focusing characterizes this relevance (Xue et al., 2023), while C^{2}KD develops transfer mechanisms that account for the modality gap (Huo et al., 2024). Alternating adaptation (Zhang et al., 2024) and gradient-aligned learning (Huang & Yang, 2026) address optimization compatibility during cross-modal interaction. Source–target alignment and compatible optimization nevertheless provide different criteria from stability under modality conditioning. PDST evaluates alternative routes with shared endpoints and uses their consensus and discrepancy to determine the transferred content and supervision strength.

3 Path-Discrepancy Guided Selective Transfer

3.1 Overview and Problem Formulation

Given a paired multimodal sample (x^A, x^B, y), modality-specific encoders extract $z_A = f_A(x^A)$ and $z_B = f_B(x^B)$, while a fusion projection constructs the multimodal representation $z_C = P([z_A; z_B])$, where C denotes the fused multimodal state rather than an additional physical modality. A shared classifier produces $p_m = \mathrm{softmax}(W z_m)$ for $m \in \{A, B, C\}$, and all modality states are jointly supervised by $\mathcal{L}_{\mathrm{cls}} = \sum_{m\in\{A,B,C\}} \mathrm{CE}(p_m, y)$ so that unimodal learning and multimodal prediction are optimized within the same task space.

Multimodal state C supplies complementary evidence for unimodal learning, but its discrepancy from a target branch mixes missing knowledge with interaction-dependent information. Uniform alignment treats both as transferable supervision. PDST separates them by comparing routes from a common source to the same target, identifying knowledge that remains invariant under modality mediation.

PDST comprises two components: *path-discrepancy modeling* identifies the consensus between alternative routes, and *discrepancy-guided selective distillation* converts the consensus into adaptively weighted supervision. Both operate in an auxiliary knowledge space during training; inference retains the original prediction pathway.

3.2 Path-Discrepancy Modeling

To model transfer without directly constraining the classification features, PDST introduces knowledge encoders E_m with $h_m = E_m(z_m) \in \mathbb{R}^{d_h}$ for $m \in \{A, B, C\}$. Task prediction continues to use the original representations z_m. Directed operators $F_{s\to t}$ map source knowledge into the target knowledge space; opposite directions use distinct operators because transfer between heterogeneous states need not be symmetric.

For three distinct states (s, r, t), the direct route is $u_{s,t} = F_{s\to t}(h_s)$ and the mediated route is $v_{s,r,t} = F_{r\to t}(F_{s\to r}(h_s))$. Holding both endpoints fixed isolates sensitivity to the intermediate modality, quantified as

$$\delta_{s,t|r} = 1 - \cos(u_{s,t}, v_{s,r,t}), \tag{1}$$

where small $\delta_{s,t|r}$ indicates stability under modality mediation. Path discrepancy characterizes transfer behavior without imposing agreement as an optimization objective, which would suppress the path-dependent variation required for selective transfer.

Geometric interpretation. A direct decomposition of alternative transfer routes provides a theoretical connection between path discrepancy and transferability. After normalizing the two route representations as $\bar{u} = u/(\|u\|_2 + \epsilon)$ and $\bar{v} = v/(\|v\|_2 + \epsilon)$ with $\delta = 1 - \bar{u}^\top \bar{v}$, their shared and route-dependent components can be written as $q^{\mathrm{inv}} = (\bar{u} + \bar{v})/2$ and $q^{\mathrm{pd}} = (\bar{u} - \bar{v})/2$, respectively. Consensus and residual components distinguish knowledge shared across routes from variation sensitive to modality mediation.

Proposition 1 (Path-discrepancy decomposition). For any pair of unit-normalized route representations $\bar{u}$ and $\bar{v}$, the corresponding route-consistent and route-dependent components satisfy

$$\left\langle q^{\mathrm{inv}}, q^{\mathrm{pd}} \right\rangle = 0, \qquad \left\| q^{\mathrm{inv}} \right\|_2^2 = 1 - \frac{\delta}{2}, \qquad \left\| q^{\mathrm{pd}} \right\|_2^2 = \frac{\delta}{2}. \tag{2}$$

Proof. Substituting the definitions of q^{inv} and q^{pd} gives $\langle q^{\mathrm{inv}}, q^{\mathrm{pd}} \rangle = \frac{1}{4}(\|\bar{u}\|_2^2 - \|\bar{v}\|_2^2) = 0$ because both route representations are unit normalized. Using $\bar{u}^\top \bar{v} = 1 - \delta$ further yields $\|q^{\mathrm{inv}}\|_2^2 = 1 - \delta/2$ and $\|q^{\mathrm{pd}}\|_2^2 = \delta/2$ after direct expansion, while minimizing $\|q - \bar{u}\|_2^2 + \|q - \bar{v}\|_2^2$ gives $q = (\bar{u} + \bar{v})/2 = q^{\mathrm{inv}}$, showing that the invariant component is simultaneously the least-squares consensus of the alternative routes. □

Proposition 1 establishes path discrepancy as an explicit measure of how representation energy is distributed between route-consistent and route-dependent components. Increasing δ transfers an equal amount of energy away from the consensus component and toward route-dependent variation, whereas decreasing δ indicates that a larger fraction of the source knowledge survives modality mediation without changing its expression. Transfer confidence can consequently be derived directly from the invariant energy as $\omega_{s,t|r} = \mathrm{sg}(1 - \delta_{s,t|r}/2)$, which naturally lies in $[0, 1]$ because $\delta_{s,t|r} \in [0, 2]$ and avoids introducing an additional temperature or manually designed nonlinear mapping. Stop-gradient isolates transferability estimation from the downstream distillation objective, preventing the transport operators from artificially increasing their transfer confidence by manipulating path discrepancy.

Accurate path characterization further requires the auxiliary knowledge space to preserve meaningful source information and the directed transport operators to avoid degenerate mappings, for which PDST employs a lightweight calibration objective without altering the discrepancy criterion itself. Reconstruction preserves the information represented by each state through $\mathcal{L}_{\mathrm{rec}} = \sum_m \|R_m(h_m) - \mathrm{sg}(z_m)\|_2^2$, while cycle consistency regularizes cross-state transport through $\mathcal{L}_{\mathrm{cyc}} = \sum_{s \neq t} \mathcal{D}_{\cos}(F_{t \to s}(F_{s \to t}(h_s)), \mathrm{sg}(h_s))$, where $\mathcal{D}_{\cos}(a, b) = 1 - \cos(a, b)$. Combining both terms as $\mathcal{L}_{\mathrm{cal}} = \mathcal{L}_{\mathrm{rec}} + \beta \mathcal{L}_{\mathrm{cyc}}$ maintains an informative auxiliary space and stable transport structure without encouraging alternative paths to collapse toward identical representations.

3.3 Discrepancy-Guided Selective Distillation

Selective distillation constructs the teacher from path-invariant knowledge and derives its supervision weight from invariant energy. Teacher construction and confidence estimation therefore share a common geometric basis.

For target modality $t \in \{A, B\}$ and the remaining physical modality r, multimodal knowledge is evaluated through the direct route $C \to t$ and the mediated route $C \to r \to t$. Their normalized representations $\bar{u}_{C,t}$ and $\bar{v}_{C,r,t}$ define the consensus teacher $q_{t|r}^{\mathrm{inv}} = \mathrm{norm}((\bar{u}_{C,t} + \bar{v}_{C,r,t})/2)$, where $\mathrm{norm}(a) = a/(\|a\|_2 + \epsilon)$, while a lightweight prediction head G_t maps the target knowledge state into the corresponding distillation space. Selective knowledge transfer is then performed through

$$\mathcal{L}_{\mathrm{sel}} = \sum_{(C,r,t) \in \mathcal{P}_{\mathrm{tr}}} \omega_{C,t|r}\, \mathcal{D}_{\cos}\Big(G_t(h_t), \mathrm{sg}\Big(q_{t|r}^{\mathrm{inv}}\Big)\Big), \tag{3}$$

where $\mathcal{P}_{\mathrm{tr}} = \{(C, B, A), (C, A, B)\}$ enhances both unimodal branches. Small path discrepancy yields stronger supervision, while larger discrepancy attenuates it as more source energy becomes route dependent.

Path-dependent variation may encode interactions unavailable to an isolated target modality. PDST therefore excludes it from the teacher while retaining it through the original multimodal task objective. Consensus construction selects the transferred content, and invariant energy determines its influence on the target branch. Selective supervision strengthens modality-specific learning without enforcing alignment with interaction-dependent information. Training combines task prediction, selective transfer, and auxiliary calibration through

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{sel}} \mathcal{L}_{\mathrm{sel}} + \lambda_{\mathrm{cal}} \mathcal{L}_{\mathrm{cal}}. \tag{4}$$

Knowledge encoders, directed transport operators, reconstruction heads, and distillation heads participate only in training, while inference retains the original modality encoders, fusion module, and classifier without introducing additional computational overhead.

For modalities $\{M_1, \ldots, M_M\}$ with joint representation $J = P([z_1; \ldots; z_M])$, compare the direct route $J \to M_i$ with mediated routes $J \to M_j \to M_i$ for $j \neq i$. Path decomposition and discrepancy-guided weighting determine the teacher and transfer strength for each target modality. Selecting a subset of intermediate states limits computational growth with the number of modalities.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate PDST on six multimodal classification benchmarks spanning audio-visual, image-text, and multi-stream settings. CREMA-D (Cao et al., 2014), Kinetics-Sounds (Arandjelovic & Zisserman, 2017), and VGGSound (Chen et al., 2020) support audio-visual recognition. Twitter2015 sentiment annotations (Yu & Jiang, 2019) and Sarcasm (Cai et al., 2019) support image-text classification. NVGesture provides RGB, optical-flow, and depth streams for evaluating the extension beyond bimodal learning (Molchanov et al., 2016). CREMA-D contains 6,698/744 training/test clips, Kinetics-Sounds contains 15,000/1,900/1,900 training/validation/test clips, and VGGSound contains 168,618 training-and-validation videos and 13,954 test videos. Twitter2015 and Sarcasm contain 3,197/1,122/1,037 and 19,816/2,410/2,409 training/validation/test pairs, respectively, while NVGesture contains 1,050 training and 482 test samples.

Implementation Details and Evaluation Metrics. Audio-visual experiments employ randomly initialized ResNet-18 backbones (He et al., 2016) for both modality streams, while NVGesture uses I3D (Carreira & Zisserman, 2017) with the same pretrained initialization as the reference implementation. Image-text experiments employ BERT (Devlin et al., 2019) and ResNet-50 as the text and image encoders, respectively. Audio-visual and NVGesture models are optimized using SGD with momentum 0.9, weight decay 10^{-4}, and an initial learning rate of 10^{-2}, which is reduced by a factor of 10 after validation performance saturates; batch sizes are 64 for CREMA-D and Kinetics-Sounds, 16 for VGGSound, and 2 for NVGesture. Image-text models are optimized using Adam (Kingma & Ba, 2015) with an initial learning rate of 2×10^{-5} and a batch size of 32. Comparisons under the common evaluation protocol use the same backbones and initialization; separately reported results are identified below. PDST introduces path-discrepancy modeling and selective distillation exclusively during training, leaving the original multimodal inference pathway unchanged. Accuracy and mean average precision (mAP) are reported for audio-visual benchmarks, while Accuracy and Macro-F1 are used for Twitter2015, Sarcasm, and NVGesture. Results of PDST are averaged over three random seeds, with $\pm$ denoting the standard deviation in percentage points.

Compared Methods. Comparisons include Concat fusion, OGM (Peng et al., 2022), AGM (Li et al., 2023), MMPareto (Wei & Hu, 2024), SMV (Wei et al., 2024a), MLA (Zhang et al., 2024), ReconBoost (Hua et al., 2024), DLMG/LFM (Yang et al., 2024), and AUG (Jiang et al., 2025b). BSS-L (Guan et al., 2025), IPRM (Yang et al., 2025b), and IGM (Jiang et al., 2025a) extend the comparison to recent balancing approaches. Results unavailable under our common evaluation protocol are taken from the original papers and marked with †.

4.2 MAIN RESULTS

Tables 1 and 2 compare PDST with fusion and balancing baselines across six benchmarks. PDST achieves the highest accuracy and mAP or Macro-F1 among the compared methods on all three audio-visual and both image-text datasets. Comparisons against Concat and the strongest baseline distinguish absolute gains from relative advantages.

Audio-visual recognition. Relative to Concat, PDST improves accuracy by 22.28, 11.10, and 2.25 percentage points on CREMA-D, Kinetics-Sounds, and VGGSound, respectively, with corresponding mAP gains of 22.85, 10.80, and 5.12 points. VGGSound exhibits a larger improvement in mAP than in accuracy, indicating that the benefit extends to class-wise prediction ranking beyond changes in top-1 classification. Against the best-performing baseline for each metric, the accuracy margins are 0.74, 1.28, and 0.40 points, while the mAP margins are 0.23, 1.47, and 0.38 points. Kinetics-Sounds exhibits the largest margins on both metrics, although CREMA-D has the largest gain over Concat. Benchmark-dependent margins distinguish gains over the fusion model from advantages over competing balancing methods. Positive margins across all three datasets support the effectiveness of selective transfer across different audio-visual recognition settings.

Image-text and multi-stream recognition. On Twitter2015, Concat underperforms the best unimodal model by 3.56 accuracy points and 4.63 F1 points, indicating that joint access to both modalities

Table 1: Audio-visual recognition results (%). Values for PDST are mean ± standard deviation over three seeds. † marks results reported in the original papers; bold and underline indicate the best and second-best scores.

	CREMA-D		Kinetics-Sounds		VGGSound	
Method	Acc.	mAP	Acc.	mAP	Acc.	mAP
Best unimodal	63.17	68.61	55.62	58.37	46.55	47.01
Concat	63.61	68.41	64.55	71.30	51.16	53.52
OGM	66.12	73.72	65.82	71.59	48.29	49.78
AGM	67.33	78.07	67.91	73.88	47.11	51.98
MMPareto	74.87	85.35	70.00	78.50	51.25	54.73
SMV	78.72	84.17	69.00	74.26	50.31	53.62
MLA	79.43	85.72	70.04	79.45	51.65	54.73
ReconBoost	75.57	81.40	68.55	76.62	50.97	53.87
DLMG/LFM	83.62	90.06	72.53	78.97	52.74	56.94
AUG	<u>85.15</u>	<u>91.03</u>	72.63	79.01	<u>53.01</u>	<u>58.26</u>
BSS-L†	82.80	88.61	73.95	79.43	52.80	56.61
IPRM†	84.27	90.66	<u>74.37</u>	<u>80.63</u>	–	–
IGM†	81.05	89.48	74.03	78.55	–	–
PDST	**85.89**±0.37	**91.26**±0.12	**75.65**±0.33	**82.10**±0.58	**53.41**±0.23	**58.64**±0.24

Table 2: Image-text and multi-stream results (%). F1 denotes Macro-F1. Values for PDST are mean ± standard deviation over three seeds. †, bold, and underline follow Table 1.

	Twitter2015		Sarcasm		NVGesture	
Method	Acc.	F1	Acc.	F1	Acc.	F1
Best unimodal	73.67	68.49	81.36	80.56	81.54	81.83
Concat	70.11	63.86	82.86	82.40	82.37	82.70
OGM	70.58	64.35	83.60	82.93	–	–
AGM	72.61	65.02	83.06	82.93	82.79	82.84
MMPareto	73.58	67.29	83.48	82.84	83.82	84.24
SMV	74.28	68.17	84.18	83.68	83.52	83.41
MLA	73.52	67.13	84.26	83.48	83.40	83.72
ReconBoost	74.42	68.32	84.37	83.17	83.86	84.34
DLMG/LFM	75.01	<u>70.57</u>	84.97	84.57	84.36	84.68
AUG	75.12	69.62	85.10	84.58	85.01	85.33
BSS-L†	<u>75.22</u>	69.51	85.01	<u>84.62</u>	<u>86.72</u>	**87.04**
IPRM†	–	–	<u>85.14</u>	84.41	85.89	86.34
IGM†	74.89	69.17	84.68	83.92	**86.93**	<u>87.03</u>
PDST	**76.28**±0.46	**71.31**±0.42	**85.47**±0.21	**84.86**±0.12	85.48±0.34	85.82±0.28

is insufficient for effective multimodal learning. PDST exceeds the best unimodal result by 2.61 accuracy points and 2.82 F1 points, improving both overall accuracy and macro-averaged predictive performance. Relative to Concat, the F1 gain of 7.45 points exceeds the accuracy gain of 6.17 points. Performance gains consequently persist under equal weighting of individual classes. Against the best-performing baselines, PDST gains 1.06 accuracy points and 0.74 F1 points. On Sarcasm, the corresponding margins are 0.33 and 0.24 points, respectively. Consistent performance ordering across both tasks supports selective transfer in image-text learning, with the magnitude of improvement depending on the task and comparison baseline.

On NVGesture, PDST improves on Concat by 3.11 accuracy points and 3.12 F1 points, and exceeds AUG by 0.47 and 0.49 points. Its 85.48% accuracy remains below the reported IGM result of 86.93%, and its 85.82% F1 is below the reported BSS-L result of 87.04%. Evaluation with three streams therefore supports an improvement over conventional fusion, although the reported IGM and BSS-L results remain stronger on their respective metrics. Comparative advantages are most consistent

in audio-visual and image-text recognition, while NVGesture demonstrates the applicability of the transfer-path formulation beyond two modalities.

4.3 Ablation Studies

Component effectiveness. Figure 2 decomposes the accuracy gain over Concat into sequential additions. Vanilla KD introduces direct multimodal distillation as the initial transfer baseline. Path-discrepancy modeling (PDM) adopts the path-invariant consensus as the distillation target under uniform weighting, and discrepancy-guided selective distillation (DSD) adds sample-wise weighting. Vanilla KD provides the largest initial gains on CREMA-D and Kinetics-Sounds, at 13.23 and 5.17 points. Replacing that target with the PDM consensus adds another 8.12 and 5.01 points. On VGGSound, Sarcasm, and NVGesture, the PDM increment is larger than the Vanilla KD increment, indicating that target selection matters even where direct distillation offers relatively little improvement.

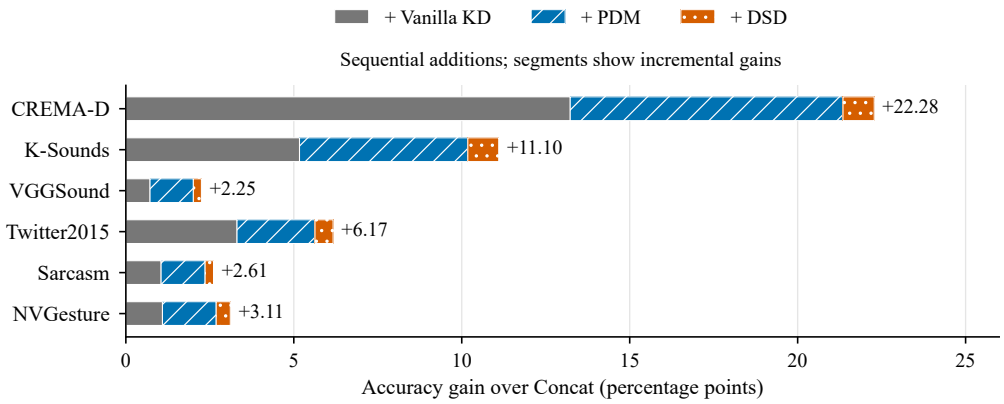


Figure 2: Sequential component contributions. Bar segments show the additional accuracy gained by Vanilla KD, PDM, and DSD in that order; labels give the total gain over Concat in percentage points. K-Sounds abbreviates Kinetics-Sounds throughout the analysis figures.

DSD adds 0.24–0.93 points across the six datasets, with a smaller increment than PDM in each case. Under sequential component addition, PDM accounts for 79.2–89.7% of the improvement obtained after Vanilla KD. Target construction accounts for the dominant increment, with confidence weighting providing an additional refinement. Positive DSD increments indicate that sample-dependent weighting remains beneficial after consensus selection. Component contributions distinguish teacher construction, which selects the transferred component, from confidence estimation, which controls its contribution to the target objective.

Transferability criterion. Figure 3(a) holds the selective-distillation framework fixed and varies the confidence criterion. Prediction discrepancy consistently improves on feature discrepancy, and both outperform uniform weighting. Path discrepancy adds a further 0.16–0.58 accuracy points over prediction discrepancy, with the largest increments on CREMA-D and Kinetics-Sounds. Endpoint-based criteria therefore contain information relevant to transfer, but they do not fully capture the reliability revealed by comparing alternative routes. Because every variant distills the same consensus, observed differences isolate the confidence criterion rather than changes in teacher content or transfer-path construction. Agreement with the component ablation supports path consistency as a basis for both teacher construction and confidence estimation.

Transfer-path construction. Direct transfer outperforms mediated transfer on all six datasets, while their joint use improves on either route (Figure 3(b)). Adding the mediated route to direct transfer yields 0.16–0.54 accuracy points; explicitly comparing the two routes yields a further 0.20–0.83 points. Explicit path comparison contributes the larger increment on every dataset. On Kinetics-Sounds, the corresponding accuracies are 74.31%, 74.82%, and 75.65%. Since joint transfer and path comparison use the same routes, access to an additional pathway does not account for the

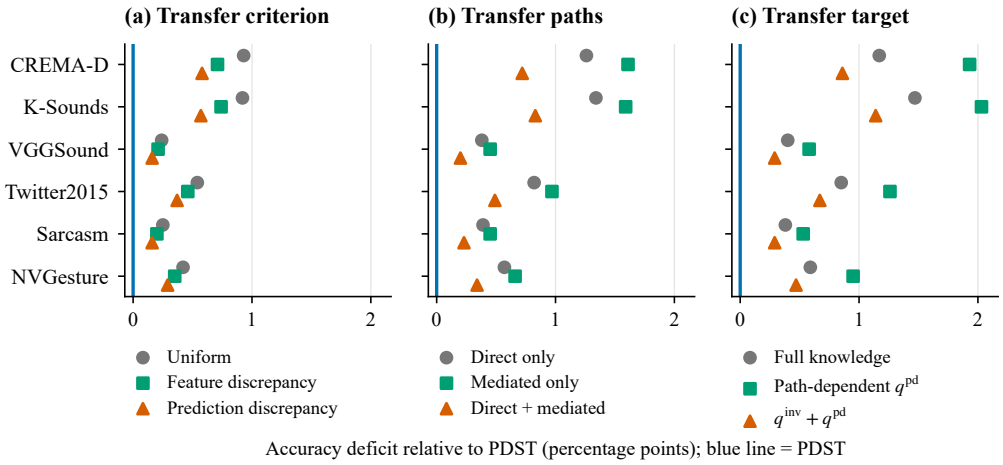


Figure 3: Design choices for (a) transfer confidence, (b) transfer paths, and (c) the distillation target. Each point is the accuracy of full PDST minus that of the indicated alternative, in percentage points. Blue vertical lines denote the full-model reference; larger deficits indicate a larger loss from replacing that choice. All panels share the same horizontal scale.

full improvement. Cross-path agreement provides an additional criterion for selecting supervision beyond the content of individual transfer routes.

Distillation target. Figure 3(c) compares full knowledge, path-dependent variation q^{pd}, the combined components $q^{\mathrm{inv}} + q^{\mathrm{pd}}$, and the consensus q^{inv}. Consensus-based supervision achieves the highest accuracy on every dataset. Reintroducing q^{pd} reduces accuracy by 0.29–1.14 points, and using it alone produces the largest deficits, at 0.53–2.03 points. On Kinetics-Sounds, the consensus, combined components, and path-dependent target yield 75.65%, 74.51%, and 73.62%, respectively. Performance ordering indicates that additional transferred content is not uniformly beneficial to the unimodal learner. Excluding variation sensitive to modality mediation produces a more effective distillation target. Path-dependent information remains available to the fused predictor through the original task objective, so the comparison concerns its suitability as unimodal supervision.

4.4 Further Analysis

Discrepancy and empirical transfer benefit. Empirical transfer benefit is defined as the reduction in target prediction loss induced by knowledge transfer. Figure 4(a) reports its Spearman correlation with each discrepancy criterion. Path discrepancy exhibits the strongest negative association on every dataset, with ρ ranging from -0.62 to -0.49, compared with -0.38 to -0.26 for prediction discrepancy and -0.31 to -0.21 for feature discrepancy. Path discrepancy therefore provides a closer rank association with transfer benefit than either endpoint measure. Consistency with the criterion ablation in Figure 3(a) connects a more informative transferability signal to higher task accuracy under the same selective-distillation framework.

Figure 4(b) further shows that transfer benefit decreases at every successive discrepancy quantile on each dataset. Relative reductions from Q_1 to Q_5 range from 76.6% on CREMA-D to 84.3% on VGGSound. Monotonic within-dataset trends agree with the decreasing confidence function $\omega = 1 - \delta/2$: samples with greater path consistency receive larger distillation weights. Nevertheless, transfer benefit remains positive in the highest-discrepancy group across all six datasets. Disagreement therefore identifies reduced supervision utility, rather than a uniformly detrimental transfer regime, supporting continuous attenuation instead of unconditional rejection of high-discrepancy samples.

Selective alignment of modality-specific representations. Figure 5 reports cosine similarity to q^{inv} and q^{pd}, averaged over target modalities and test samples. Vanilla KD increases similarity to both components relative to the baseline, including a 0.12–0.15 increase toward q^{pd}. PDST instead raises similarity to q^{inv} by another 0.10–0.15 over Vanilla KD while reducing similarity to q^{pd} by

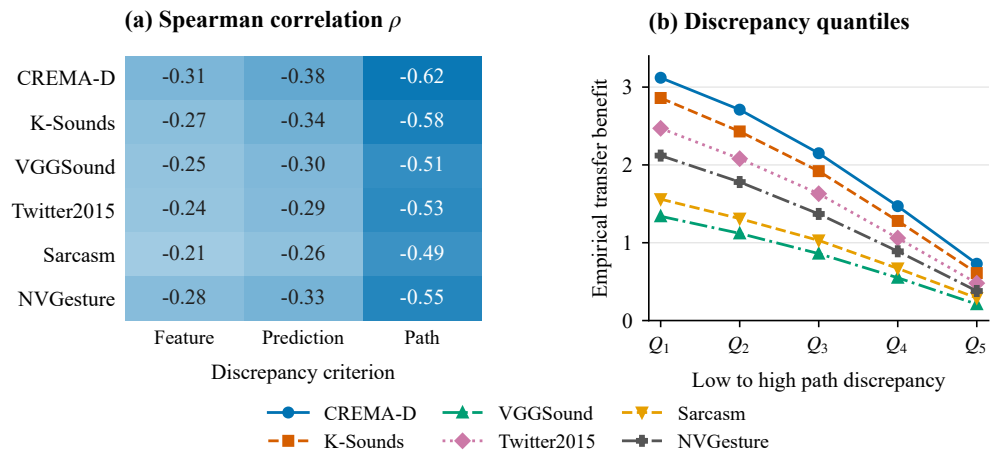


Figure 4: Discrepancy and empirical transfer benefit. (a) Signed Spearman correlations, with darker cells denoting a stronger negative association. (b) Transfer benefit on its reported loss-reduction scale across path-discrepancy quantiles, from the lowest (Q_1) to the highest (Q_5). Each curve represents one dataset; markers and line styles distinguish the curves independently of color.

0.11–0.13. Its path-dependent similarity remains only 0.01–0.02 above the baseline. Consequently, the difference between invariant and path-dependent similarities increases from 0.17–0.27 under Vanilla KD to 0.39–0.52 under PDST. Increased separation characterizes selective representation alignment rather than indiscriminate similarity to transferred knowledge. Agreement with the transfer-target ablation supports an association between improved task performance and preferential alignment with path-invariant knowledge.

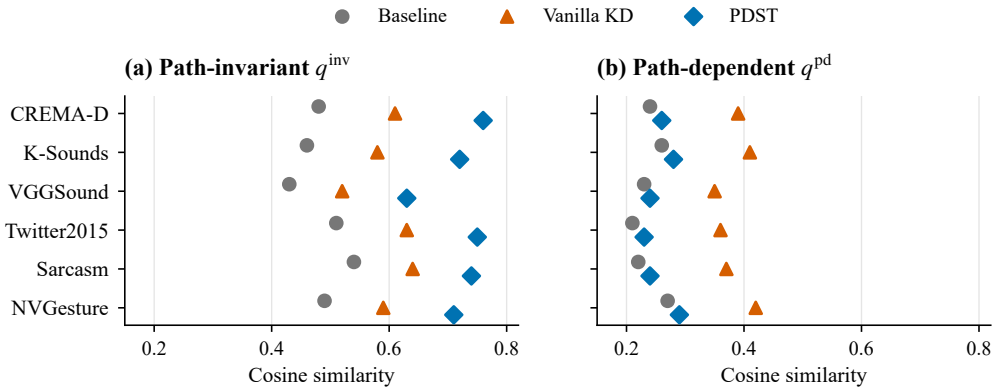


Figure 5: Cosine similarity of modality-specific representations to (a) path-invariant and (b) path-dependent knowledge. Both panels use the same scale. Compared with Vanilla KD, PDST aligns more strongly with the consensus while retaining near-baseline similarity to path-dependent variation.

5 Conclusion

PDST formulates selective multimodal distillation through consistency between direct and mediated transfer routes. Path-invariant consensus defines the teacher, and invariant energy determines the transfer weight, while path-dependent variation remains available to the multimodal task. Across six benchmarks, PDST improves on conventional fusion and achieves its strongest comparative results in audio-visual and image-text learning. Ablations distinguish the benefits of consensus selection and adaptive weighting, while correlation and representation analyses connect path discrepancy to transfer benefit and selective alignment. Empirical evidence supports transfer-path consistency as a criterion for selecting distillation targets and weighting supervision without modifying the inference model.

AI Assistance Disclosure

The authors used generative AI tools for language editing and proofreading to improve the clarity and readability of the manuscript. All scientific content, including ideas, methods, experiments, and conclusions, was generated and verified by the authors.

References

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, pp. 609–617, 2017. doi: 10.1109/ICCV.2017.73. URL https://dblp.org/rec/conf/iccv/ArandjelovicZ17.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *ACL*, pp. 2506–2515, 2019. doi: 10.18653/v1/P19-1239. URL https://dblp.org/rec/conf/acl/CaiCW19.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244. URL https://dblp.org/rec/journals/taffco/CaoCKGNV14.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pp. 4724–4733, 2017. URL https://dblp.org/rec/conf/cvpr/CarreiraZ17.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *ICASSP*, pp. 721–725, 2020. doi: 10.1109/ICASSP40776.2020.9053174. URL https://dblp.org/rec/conf/icassp/ChenXVZ20.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423. URL https://dblp.org/rec/conf/naacl/DevlinCLT19.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *ICML*, pp. 8632–8656, 2023. URL https://dblp.org/rec/conf/icml/DuTLLYWYZ23.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: Prototypical modal rebalance for multimodal learning. In *CVPR*, pp. 20029–20038, 2023. doi: 10.1109/CVPR52729.2023.01918. URL https://dblp.org/rec/conf/cvpr/Fan0WW023.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. doi: 10.1007/s11263-021-01453-z. URL https://dblp.org/rec/journals/ijcv/GouYMT21.

Zhi-Hao Guan, Qing-Yuan Jiang, and Yang Yang. Balance-aware sequence sampling makes multi-modal learning better. In *IJCAI*, pp. 2838–2846, 2025. URL https://dblp.org/rec/conf/ijcai/GuanJY25.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *CVPR*, pp. 2827–2836, 2016. doi: 10.1109/CVPR.2016.309. URL https://dblp.org/rec/conf/cvpr/GuptaHM16.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://dblp.org/rec/conf/cvpr/HeZRS16.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL https://dblp.org/rec/journals/corr/HintonVD15.

Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. Reconboost: Boosting can achieve modality reconcilement. In *ICML*, pp. 19573–19597, 2024. URL https://dblp.org/rec/conf/icml/HuaXB0H24.

Longfei Huang and Yang Yang. Harmonized tabular-image fusion via gradient-aligned alternating learning. *CoRR*, abs/2604.01579, 2026. doi: 10.48550/arXiv.2604.01579. URL https://dblp.org/rec/journals/corr/abs-2604-01579.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably). In *ICML*, pp. 9226–9259, 2022. URL https://dblp.org/rec/conf/icml/HuangLZYH22.

Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. C2KD: Bridging the modality gap for cross-modal knowledge distillation. In *CVPR*, pp. 16006–16015, 2024. doi: 10.1109/CVPR52733.2024.01515. URL https://dblp.org/rec/conf/cvpr/Huo0GW024.

Qing-Yuan Jiang, Zhouyang Chi, and Yang Yang. Interactive multimodal learning via flat gradient modification. In *IJCAI*, pp. 5489–5497, 2025a. URL https://dblp.org/rec/conf/ijcai/JiangCY25.

Qing-Yuan Jiang, Longfei Huang, and Yang Yang. Rethinking multimodal learning from the perspective of mitigating classification ability disproportion. In *NeurIPS*, 2025b. URL https://dblp.org/rec/conf/nips/JiangHY25.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. URL https://dblp.org/rec/journals/corr/KingmaB14.

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *ICCV*, pp. 22157–22167, 2023. URL https://dblp.org/rec/conf/iccv/LiLHLLZ23.

Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. In *CVPR*, pp. 4207–4215, 2016. URL https://dblp.org/rec/conf/cvpr/MolchanovYGKTK16.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *ICML*, pp. 689–696, 2011. URL https://dblp.org/rec/conf/icml/NgiamKKNLN11.

Hongpeng Pan and Yang Yang. Coordinated uni-modal assistance for enhancing multi-modal learning. In *ICME*, pp. 1–6, 2025. URL https://dblp.org/rec/conf/icmcs/PanY25.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, pp. 8228–8237, 2022. doi: 10.1109/CVPR52688.2022.00806. URL https://dblp.org/rec/conf/cvpr/PengWD0H22.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, pp. 12692–12702, 2020. doi: 10.1109/CVPR42600.2020.01271. URL https://dblp.org/rec/conf/cvpr/WangTF20.

Yake Wei and Di Hu. MMPareto: Boosting multimodal learning with innocent unimodal assistance. In *ICML*, pp. 52559–52572, 2024. URL https://dblp.org/rec/conf/icml/Wei024.

Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *CVPR*, pp. 27328–27337, 2024a. doi: 10.1109/CVPR52733.2024.02581. URL https://dblp.org/rec/conf/cvpr/WeiF0024.

Yake Wei, Siwei Li, Ruoxuan Feng, and Di Hu. Diagnosing and re-learning for balanced multimodal learning. In *ECCV*, pp. 71–86, 2024b. doi: 10.1007/978-3-031-73039-9_5. URL https://dblp.org/rec/conf/eccv/WeiLFH24.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *ICML*, pp. 24043–24055, 2022. URL https://dblp.org/rec/conf/icml/WuJCG22.

Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. In *ICLR*, 2023. URL https://dblp.org/rec/conf/iclr/XueGRZ23.

Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating multimodal classification via dynamically learning modality gap. In *NeurIPS*, 2024. URL https://dblp.org/rec/conf/nips/0074WJ024.

Yang Yang, Hongpeng Pan, Qing-Yuan Jiang, Yi Xu, and Jinhui Tang. Learning to rebalance multi-modal optimization by adaptively masking subnetworks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4553–4566, 2025a. doi: 10.1109/TPAMI.2025.3547417. URL https://dblp.org/rec/journals/pami/YangPJXT25.

Yang Yang, Xixian Wu, and Qing-Yuan Jiang. Towards equilibrium: An instantaneous probe-and-rebalance multimodal learning approach. In *IJCAI*, pp. 3552–3560, 2025b. doi: 10.24963/IJCAI.2025/395. URL https://dblp.org/rec/conf/ijcai/YangWJ25.

Feng Yu, Xiangyu Wu, Yang Yang, and Jianfeng Lu. Multimodal classification via total correlation maximization. *CoRR*, abs/2602.13015, 2026. doi: 10.48550/arXiv.2602.13015. URL https://dblp.org/rec/journals/corr/abs-2602-13015.

Jianfei Yu and Jing Jiang. Adapting BERT for target-oriented multimodal sentiment classification. In *IJCAI*, pp. 5408–5414, 2019. URL https://dblp.org/rec/conf/ijcai/Yu019.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *CVPR*, pp. 27446–27456, 2024. URL https://dblp.org/rec/conf/cvpr/0006YBY24.