MODfinity: Unsupervised Domain Adaptation with Multimodal Information Flow Intertwining

Supplementary Material

1. Appendix

1.1. More Details About Datasets

Our evaluation employs three datasets: AVE [8], EPIC-Kitchens 55 [2], and CogBeacon [7]. These datasets feature distinct characteristics and cover various domains, including video event detection, action recognition, and environmental interaction understanding.

1.1.1. Event Recognition Dataset (AVE).

AVE [8] dataset contains 4,143 videos covering 28 event categories, and the image modal and audio modal are aligned in time. The AVE dataset covers a wide range of audio-visual events (e.g., man speaking, dog barking, playing guitar, frying food, etc.), and each video contains at least one 2s long audio-visual event. We follow [5] to gain the source domain and the target domain. Specifically, the Resnet-50 network [3] pre-trained on Imagenet is used to extract 1024-dimensional features of the image of each sample. Then the feature vectors of each category are clustered into two clusters by the K-Means algorithm. Following [5], we obtained 41,728 source domain samples and 23,919 target domain samples. Examples of 12 different categories of images in the source domain and the target domain are shown in Fig 2. It can be seen from Fig 2 that the pictures of the source domain are distinguishable, while images in the target domain are more difficult to distinguish due to poor lighting conditions or occlusion. This shows the obvious domain shift between the source domain and the target domain.

1.1.2. Action Recognition Dataset (EPIC-Kitchens 55).

EPIC-Kitchens 55 [2] is a multi-modal dataset designed to test domain adaptation for action recognition. It is recorded in 32 environments and contains two modal forms of RGB image and Optical Flow. In this paper, we use the same division method as the previous work [6], considering the domain adaptation problem among the three domains D1, D2, and D3 in EPIC Kitchens. Some scenes of image and optical flow models in the three domains are shown in Fig 1, which reflects the shift between domains. Eight types of actions are analyzed: ('put', 'take', 'open', 'close', 'wash', 'cut', 'mix', and 'pour'). The number of action segments in the three domains D1, D2, and D3 are 1,978, 3,245 and 4,871 respectively. Even though this ensures sufficient examples per domain and class, EPIC-Kitchens 55 has a large class imbalance offering additional challenges for domain adaptation. Six domain migration combinations can be obtained by combining different domains.

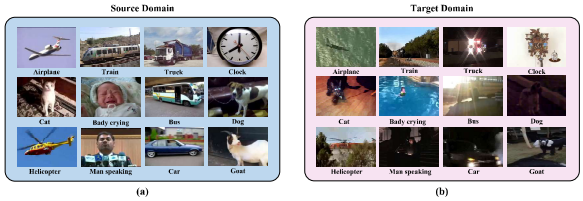


Figure 1. Examples of the instances in the domain D1, D2, and D3 in EPIC Kitchens 55.

1.1.3. Fatigue Detection Dataset (CogBeacon).

CogBeacon [7] is a multi-modal dataset designed to research the effects of cognitive fatigue on human performance. The dataset consists of 76 sessions collected from 19 male and female users performing different versions of a cognitive task inspired by the principles of the Wisconsin Card Sorting Test. During each session, the users' EEG functionality and facial key points are recorded and labeled. Specifically, each user performed three versions (namely V1, V2, and V3) of cognitive task tests. Different versions of cognitive tasks will produce different stimuli for users. For example, the EEG signals when facing text-based stimuli and sound-based stimuli are different. Therefore, the data collected under different versions of cognitive tasks can be regarded as cross-domain data. In this paper, we choose one version of cognitive task as the source domain, and the other version as the target domain. The number of samples corresponding to the cognitive tasks V1, V2, and V3 are 2,259, 2,221, and 2,389 respectively. In the experiment, we regard one of the domains as the source domain and the other domain as the target domain. This setting method can obtain six domain migration combinations.

1.2. More Details About Implementation

Our experiments are conducted using PyTorch on three NVIDIA GeForce RTX 3090 GPUs. All of our results were conducted in five or more experiments.

For the AVE dataset, RGB images are resized and center-cropped to 224×224 pixels, while audio is transformed into spectrograms using the Short-Time Fourier Transform. The feature encoders for both image and audio utilize ResNet-18 [3], transforming input samples into 512-dimensional feature vectors. Each feature encoder was followed by a two-layer fully connected head. The affinity measurement encoder consisted of three hidden layers with 512-

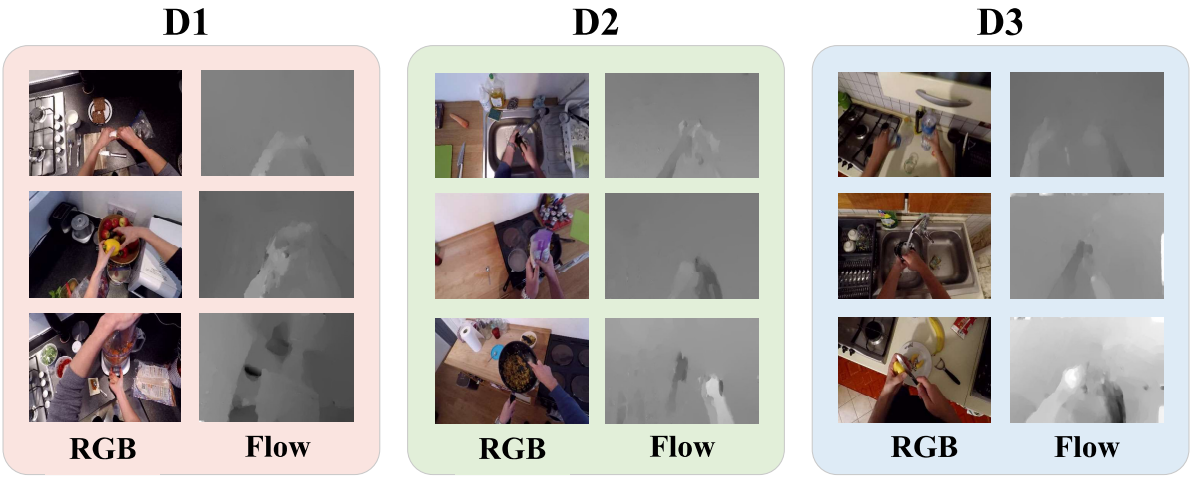


Figure 2. Examples of the images in AVE. (a) Examples in the source domain. (b) Examples in the target domain.

dimensional fully connected layers. We used an SGD optimizer with a learning rate of 0.001 and a batch size of 128 for optimization, training for 40 epochs on the source domain data. In the domain adaptation phase, we trained for 5 epochs per iteration with a learning rate of 0.001.

For the EPIC-Kitchens 55 dataset, we conduct preprocessing operations including rotation and translation. For each video segment, we uniformly select 16 frames to construct the input tensor. We employ a dual-stream I3D network [1] as the feature encoder. Following the method described in [1], we pre-train this network on the Kinetics dataset. The affinity measurement encoder consisted of three hidden layers with 512-dimensional fully connected layers. We used an SGD optimizer with a learning rate of 0.001 and a batch size of 128 for optimization, training for 80 epochs on the source domain data. In the domain adaptation phase, we trained for 5 epochs per iteration with a learning rate of 0.001.

For the CogBeacon dataset, we follow the methodology described in [5] to convert EEG signals and facial key points into one-dimensional feature inputs. We utilize a three-layer one-dimensional ResNet network [3] as the feature encoder the EEG and facial key points modalities. The affinity measurement encoder consisted of three hidden layers with 256-dimensional fully connected layers. We used an SGD optimizer with a learning rate of 0.001 and a batch size of 128 for optimization, training for 40 epochs on the source domain data. In the domain adaptation phase, we trained for 5 epochs per iteration with a learning rate of 0.001.

1.3. Supplementary Explanation of Teaser

Figure 3 illustrates the performance of different information flow optimization methods (source-only, coarse-grained [4], and our proposed method) on the AVE dataset. We present the classification results for two *Cat* samples and three *Truck* samples from the target domain using each method. Specifically, samples labeled 1 and 3 have the ground truth label *Cat*, while samples labeled 2, 4, and 5 have the ground truth label *Truck*. The results are displayed as bar charts comprising 28 bars corresponding to the 28 categories in the AVE dataset: *Church bell*, *Male speech, man speaking*, *Bark*, *Fixed-wing aircraft, airplane*, *Race car, auto racing*, *Female speech, woman speaking*, *Helicopter*, *Violin, fiddle*, *Flute*, *Ukulele*, *Frying (food)*, *Truck*, *Shofar*, *Motorcycle*, *Acoustic guitar*, *Train horn*, *Clock*, *Banjo*, *Goat*, *Baby cry, infant cry*, *Bus*, *Chainsaw*, *Cat*, *Horse*, *Toilet flush*, *Rodents, rats, mice*, *Accordion*, and *Mandolin*. In the bar charts, the green bars represent the categories corresponding to the ground truth labels (the 12th bar represents *Truck*, and the 23rd bar represents *Cat*).

Figure 3 shows that, in the source-only model, samples 1 and 4 are classified correctly in the image modality, while samples 2 and 5 are classified correctly in the audio modality. The coarse-grained method allows sub-modalities to exchange information indiscriminately, causing each sub-modality to be contaminated by incorrect information, which degrades the model's performance. In contrast, our method reduces the transmission of incorrect information by optimizing the information flow. By finely selecting high-quality information flows, sub-modalities learn from each other effectively, promoting mutual complementarity among modalities and leading to overall improvement.

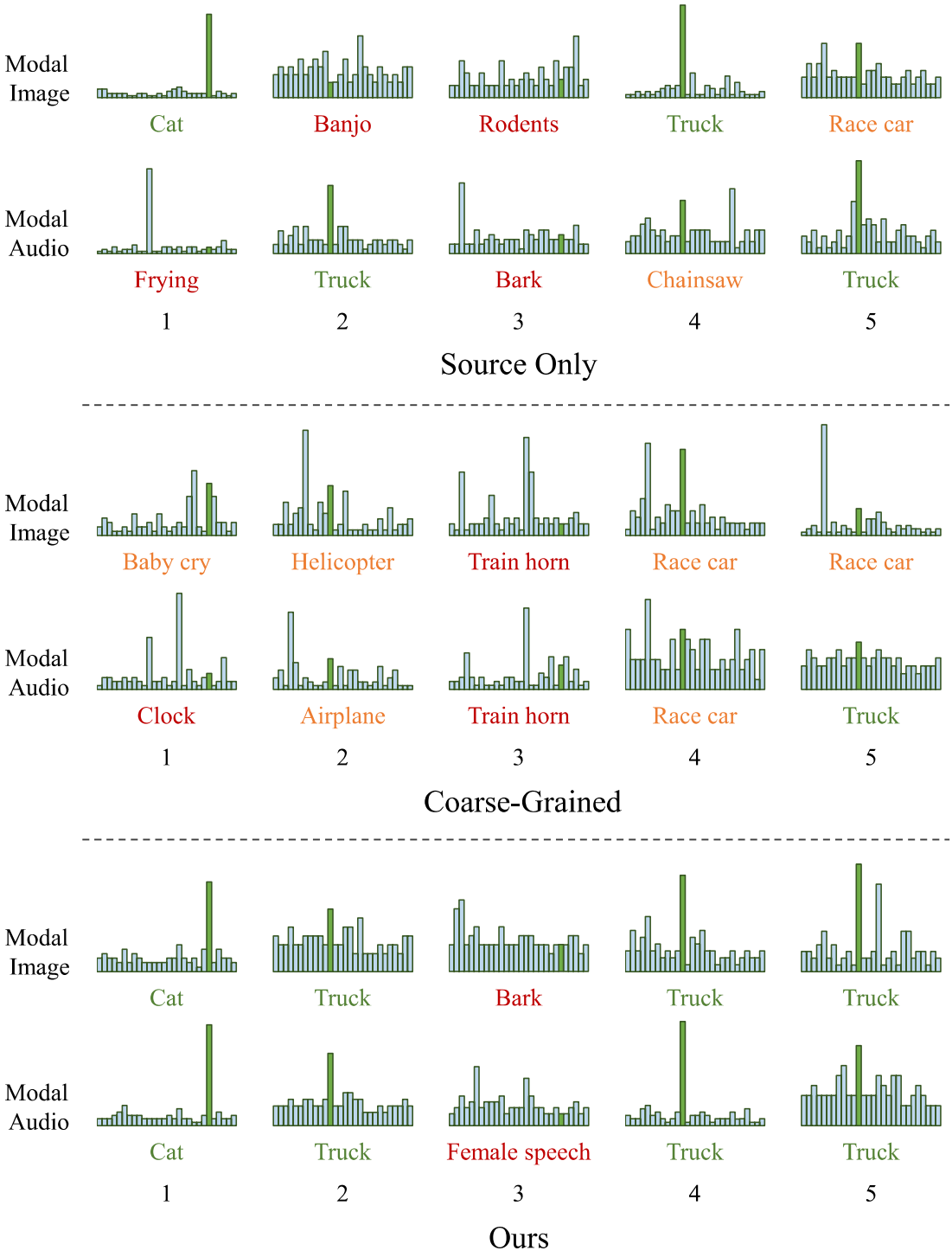


Figure 3. Detailed Description of the Teaser.This figure presents the classification vectors and the final predicted categories generated by various methods. Predictions achieving Top-1 accuracy are classified as High Performance, those falling within Top-5 accuracy are categorized as Moderate Performance, and the rest are labeled as Low Performance.

1.4. Supplementary Explanation of Efficiency issue

Our approach does not process samples sequentially; instead, we utilize batch-wise matrix operations to accelerate computation. As shown in Table 1, the training time is comparable to other related methods. Importantly, the most computationally expensive step remains backpropagation. To address this, we introduce a *gradient-blocking matrix*, which masks low-quality information flows during training. By allowing only high-quality signals to participate in backpropagation, this mechanism enhances the quality of information flow while simultaneously reducing the computational burden.

Table 1. Training Timing Cost (Second/Batch)

Method	AVE	EPIC-kitchens 55	CogBeacon	VGGSound
CE only	0.2433	3.540	0.068	4.110
MCT[25]	0.2457	3.580	0.069	4.080
DANN[9]	0.2494	3.580	0.068	4.130
CL[13]	**0.2538**	**4.200**	**0.073**	**4.540**
Ours	0.2519	3.840	0.072	4.270

We further added a time complexity analysis to address the comments. Let the batch size be B, and let the backward propagation timing costs for E^c and E^m be d_c and d_m, respectively. The time complexity at each key stage is:

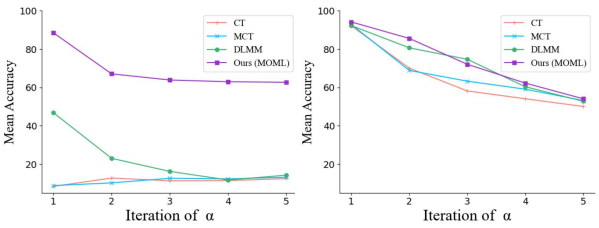
Training of Affinity Measurement. In this phase, E^c is trained using the MOML loss, while E^m is trained using the CE loss. The time complexities are $O\big(B^2 \times d_c\big)$ and $O\big(B \times d_m\big)$, respectively. Thus, the total complexity is $O\big(B^2 \times d_c + B \times d_m\big)$. Since E^m has significantly more parameters than E^c and $d_c \ll d_m$, the overall time complexity for smaller batches is approximately $O\big(B \times d_m\big)$.

Fine-Grained Sample Filtering. In this phase, each sample in the batch is compared with class feature vectors to compute an affinity matrix, with a complexity of $O\big(B \times c\big)$, where c is the number of classes. As this step involves no backpropagation, it is highly efficient. Furthermore, it is executed only when the hyperparameter α exceeds a predefined threshold, contributing to less than 5% of the total training time.

Modal-Affinity Distillation. This phase uses a gradient-blocking matrix to enable sample-level distillation via batch processing. E^c is updated with the MOML loss, while E^m is optimized using the CE and MOD losses. The total time complexity is $O\big(B^2 \times d_m\big)$, comparable to other baselines such as CL [4].

1.5. Supplementary Effectiveness Analysis of Gradient Blocking and Visualized Experiments on Noisy Data

The performance of each modality in multimodal models varies across target domains, and unconstrained information flow can amplify error propagation between modalities. Figure 6 in the original paper demonstrates that our



(a) Image Modality (b) Audio Modality

Figure 4. Accuracy of Measurements for Image and Audio Modalities in the noisy AVE Dataset.

Affinity Measurement effectively identifies high-quality information flows with high accuracy. To further validate its robustness, we include the new Figure 4, showcasing its effectiveness on noisy datasets. Additionally, in the **Fine-Grained Sample Filtering** phase, the **gradient-blocking matrix** further regulates information flows, minimizing error propagation. As highlighted in Table 4 of the original paper, our model achieves superior performance on noisy datasets, underscoring the effectiveness of these strategies.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 2

[2] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, pages 720–736, 2018. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 2

[4] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *ICCV*, pages 13618–13627, 2021. 2, 4

[5] Jianming Lv, Kaijie Liu, and Shengfeng He. Differentiated learning for multi-modal domain adaptation. In *ACM MM*, pages 1322–1330, 2021. 1, 2

[6] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*, pages 122–132, 2020. 1

[7] Michalis Papakostas, Akilesh Rajavenkatanarayanan, and Fillia Makedon. Cogbeacon: A multi-modal dataset and data-collection platform for modeling cognitive fatigue. *Technologies*, 7(2):46, 2019. 1

[8] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, pages 247–263, 2018. 1