Supplementary Materials: Defending Multimodal Fusion Models against Single-Source Adversaries

Karren Yang[1] Wan-Yi Lin[2] Manash Barman[3] Filipe Condessa[2] Zico Kolter[2,3]
[1]Massachusetts Institute of Technology [2]Bosch Center for AI [3]Carnegie Mellon University

A. Experimental Details

A.1. Model Architecture and Training

Action Recognition on Epic-Kitchens. We adapt the state-of-the-art temporal binding networks by Kazakos et al. [3]. This model extracts features from each of the three modalities using an Inception network with Batch Normalization [12]. The features from different modalities are ensembled using feed-forward networks with ReLU activation prior to temporal aggregation and prediction of a verb and noun class. We obtain a pretrained model on clean data following the data preprocessing, augmentation, and training steps of Kazakos et al. [3]. We then augment the model with the robust fusion strategy described in Section 3 of the main text and perform training according to Algorithm 1 using stochastic gradient descent with momentum 0.9 and weight decay 5×10^{-4} for $\sim$120 epochs with a learning rate of 1×10^{-3}.

Object Detection on KITTI. We adapt a single-shot detector based on YOLOv2 [9] and YOLOv3 [10] to the multimodal setting. The model extracts features from each of the three modalities using a Darknet19 network [9]. The features are fused using a 1×1 convolutional layer prior to predicting bounding boxes and confidences using a YOLO detector layer [10]. We obtain the Velodyne depth map by projecting the Velodyne points to the 2D image plane using the calibration and projection matrices provided in the dataset [2], followed by dilation with a 5×5 diamond kernel and bilateral smoothing. We obtain the stereo depth image using the pretrained PSMNet from [14]. All visual inputs are resized to 1280×380 for training and evaluation. During training, we augment the visual inputs by adding a random horizontal flip and a random shift that is at most one-fifth of the original width and height of the image, and we additionally augment the RGB image by randomly shifting the hue, exposure, and saturation. We obtain a pretrained multimodal model on unperturbed data in two steps: (i) we pretrain single-shot detectors for each modality using the same architecture, except for the fusion layer using stochastic gradient descent with momentum 0.9 and weight decay 5×10^{-4} for $\sim$500 epochs with a learning rate of 1×10^{-3}, (ii) we train the fusion layer over the pretrained Darknet19 networks with the same optimization for $\sim$120 epochs. We then augment the model with the robust fusion strategy described in Section 3 of the main text and perform training according to Algorithm 1 using stochastic gradient descent (SGD) with momentum 0.9, weight decay 5×10^{-4}, and gradient clip of 20 for $\sim$20 epochs with a learning rate of 1×10^{-3}.

Sentiment Analysis on CMU-MOSI. For video input, we use VGGFace2 [1] followed by one layer of multi-head attention [13] and two layers of bi-directional LSTMs; for audio, we apply the same architecture on the Mel-frequency cepstral coefficients (MFCCs) of the audio signal instead of VGGFace2 outputs; for text, we apply the Transformer model [13] on a 300 dimensional pretrained GloVe embedding trained on Wikipedia from [8]. We first train unimodal models using each of these three feature extractor followed by a fully-connected layer. We then augment the model with the robust fusion strategy described in Section 3 of the main text and perform training according to Algorithm 1 using stochastic gradient descent with momentum 0.9, weight decay 5×10^{-4}, a learning rate of 1×10^{-5}, and trained for 40 epochs.

A.2. Adversarial Perturbations

The adversarial perturbations considered in our experiments are white box adaptive attack, *i.e.*, attacks are generated with full knowledge of f_{robust}.

Action Recognition on EPIC-Kitchens. We consider single-source ℓ_∞ attacks on each of three modalities: vision, motion, and audio. To generate the perturbations, we approximate the solution of Equation 1 of the main text using projected gradient descent (PGD) [7], taking $\mathcal{L}$ to be the sum of the cross-entropy losses from the verb and noun classes. For vision and motion, we approximate the solution with 10-step PGD with $\epsilon = 8/256$. For the audio spectrogram, we approximate the solution with 10-step PGD with $\epsilon = 0.8$.

Object Detection on KITTI. We consider single-source ℓ_∞ perturbations on each of three visual modalities. To generate the perturbations, we approximate the solution

Odd-one-out network	**Clean**			**Visual Perturbation**			**Motion Perturbation**			**Audio Perturbation**		
	Verb	**Noun**	**Action**	**Verb**	**Noun**	**Action**	**Verb**	**Noun**	**Action**	**Verb**	**Noun**	**Action**
Random	59.6	**43.1**	30.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Unaligned features	**61.5**	42.5	**31.4**	**48.0**	**24.2**	**16.8**	**48.5**	**35.6**	**21.1**	**46.5**	**33.3**	**22.1**
Aligned features (Logits)	**61.5**	38.0	31.2	38.7	19.5	13.1	37.0	29.1	16.7	31.1	23.9	13.8

Table 1. We show the top-1 classification accuracy of our fusion model with different odd-one-out networks on action recognition on the EPIC-Kitchens dataset under clean data and single-source adversarial perturbations on each modality. Higher is better.

Odd-one-out network	**Clean**			**Visual (RGB) Perturbation**			**Depth (Velo) Perturbation**			**Stereo Disparity Perturbation**		
	Car	**Pedest.**	**Cyclist**	**Car**	**Pedest.**	**Cyclist**	**Car**	**Pedest.**	**Cyclist**	**Car**	**Pedest.**	**Cyclist**
Random	90.1	79.0	84.3	17.9	7.6	7.6	1.8	19.8	28.1	1.1	18.4	30.2
Unaligned features	**90.6**	**79.9**	**85.4**	**85.1**	**73.9**	**82.3**	**87.8**	**71.1**	**85.8**	**89.8**	**76.8**	**84.7**
Aligned features (Bounding Boxes)	90.2	76.1	83.3	77.6	66.5	73.8	83.1	64.8	68.3	80.0	65.1	44.9

Table 2. We show the average precision (at medium difficulty) of our fusion model with different odd-one-out networks on object detection on the KITTI dataset under clean data and single-source adversarial perturbations on each modality. Higher is better.

Odd-one-out network	**Clean**		**Audio Perturbation**		**Video Perturbation**		**Text Perturbation**	
	2-class	**7-class**	**2-class**	**7-class**	**2-class**	**7-class**	**2-class**	**7-class**
Random	75.41	46.73	74.48	33.52.10	61.65	31.46	62.98	25.01
Unaligned features	**82.03**	**50.89**	**73.18**	**42.06**	**69.94**	**38.20**	**66.13**	**30.20**
Aligned features	80.13	48.26	71.95	39.74	66.47	35.43	60.74	26.57

Table 3. We show the binary and 7-class accuracy of our fusion model with different odd-one-out networks on sentiment analysis on the CMU-MOSI dataset under clean data and single-source adversarial perturbations on each modality. Higher is better.

of Equation 1 in the main text using 10-step PGD with $\epsilon = 16/256$, taking $\mathcal{L}$ to be the YOLO loss [10]. The loss consists of the cross-entropy losses of object and class confidences as well as the mean-squared localization errors summed over bounding box proposals.

Sentiment Analysis on CMU-MOSI. We consider single-source attacks on each of three modalities: vision, audio, and text. The adversarial perturbations approximate the solution of Equation 1 in the main text, taking $\mathcal{L}$ to be the binary or 7-class cross entropy. For vision, we generate ℓ_∞ perturbations with 10-step PGD with $\epsilon = 8/256$. For audio, we generate the single-source ℓ_∞ perturbations with 10-step PGD with $\epsilon = 0.8$. For text, we adapted the word-replacement attack with priority based on word saliency [11] with an attack budget of one word per sentence.

A.3. Evaluation Metrics

For each evaluation metric, we consider clean performance (*i.e.*, performance on unperturbed data) as well as robust performance on data with single-source adversarial perturbations. For robust performance, we report the evaluation metric once for each attacked modality.

Action Recognition on EPIC-Kitchens. Models are evaluated based on their top-1 and top-5 classification accuracy on the verb, noun, and action classes.

Object Detection on KITTI. We consider Average Precision (AP) on three object classes: cars, pedestrians, and cyclists. For the car class, true positives are bounding boxes with Intersection over Union (IoU) > 0.7 with real boxes. For the pedestrian and cyclist classes, true positives are bounding boxes with IoU > 0.5 with real boxes.

Sentiment Analysis on CMU-MOSI. Models are evaluated based on their binary classification accuracy (*i.e.*, predicting whether the sentiment is positive or negative) as well as their classification accuracy of 7 sentiment classes.

B. Robust performance with different odd-one-out networks

In Table 5 of the main text, we show that the odd-one-out network based on unaligned representations of different features is more effective at detecting the perturbed modality than a random baseline as well as an odd-one-out network based on aligned representations of the features. Supplementary Tables 1, 2 and 3 show task performance of the models using different odd-one-out networks. Overall, we observe that task performance reflects the detection rates; when the model's odd-one-out network is more effective at detecting the perturbed modality, it is more successful at only allowing information from the unperturbed modalities to pass through the feature fusion step, resulting in better task performance under single-source perturbations. Specifically, we show that the model with the odd-one-out network based on unaligned representations of different features is more robust than the baselines that use a random odd-one-out network or an odd-one-out network based on aligned representations of the features.

Fusion	**Clean**			**Visual Perturbation**			**Motion Perturbation**			**Audio Perturbation**		
	Verb	**Noun**	**Action**	**Verb**	**Noun**	**Action**	**Verb**	**Noun**	**Action**	**Verb**	**Noun**	**Action**
Oracle (Upper Bound)	-	-	-	55.8	31.4	21.9	50.0	37.2	23.8	53.9	39.2	25.6
Concat Fusion	59.0	42.1	30.2	1.3	0.8	0.3	0.5	1.7	0.1	1.5	2.1	0.4
Mean Fusion	56.8	40.4	27.6	2.3	0.8	0.4	0.4	1.3	0.1	1.9	2.5	0.5
LEL+Robust [6]	61.2	**43.1**	30.5	54.5	30.4	21.1	51.4	36.6	22.7	51.5	37.3	23.9
Gating+Robust [5, 4]	60.9	43.0	30.6	56.9	32.5	22.7	52.0	**39.9**	25.2	55.6	39.4	26.3
Ours	**61.5**	42.5	**31.4**	**58.0**	**33.9**	**24.5**	**53.2**	39.2	**25.6**	**56.6**	**39.5**	**27.1**
Δ-Clean	**2.5**	**0.3**	**1.2**	**55.7**	**33.1**	**24.1**	**52.7**	**37.5**	**25.5**	**54.7**	**37.0**	**26.6**
Δ-Robust	**0.3**	-0.6	**0.8**	**1.1**	**1.4**	**1.8**	**1.2**	-0.7	**0.4**	**1.0**	**0.1**	**0.8**

Table 4. **Unimodal transfer attack.** Top-1 classification accuracy results on EPIC-Kitchens dataset under clean data and single-source adversarial perturbations on each modality, where the adversarial perturbations are generated and transferred from unimodal models. Higher is better.

Fusion	**Clean**			**Visual Perturbation**			**Motion Perturbation**			**Audio Perturbation**		
	Verb	**Noun**	**Action**	**Verb**	**Noun**	**Action**	**Verb**	**Noun**	**Action**	**Verb**	**Noun**	**Action**
Oracle (Upper Bound)	-	-	-	55.8	31.4	21.9	50.0	37.2	23.8	53.9	39.2	25.6
Concat Fusion	59.0	42.1	30.2	0.0	0.0	0.0	0.3	0.1	0.0	0.0	0.0	0.0
Mean Fusion	56.8	40.4	27.6	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
LEL+Robust [6]	55.8	36.7	23.9	9.4	7.2	3.1	11.3	17.8	5.1	6.2	12.3	3.1
Gating+Robust [5, 4]	57.1	39.1	26.6	18.8	12.6	6.7	32.3	25.7	14.2	16.3	13.0	7.6
Ours	**60.2**	**42.5**	**31.0**	**53.6**	**29.7**	**20.8**	**48.7**	**35.6**	**22.7**	**50.7**	**36.4**	**23.5**
Δ-Clean	**1.2**	**0.4**	**0.8**	**53.5**	**29.7**	**20.8**	**48.4**	**35.5**	**22.7**	**50.7**	**36.4**	**23.5**
Δ-Robust	**3.1**	**3.4**	**4.4**	**34.8**	**17.1**	**14.1**	**16.4**	**9.9**	**8.5**	34.4	**23.4**	**15.9**

Table 5. **Feature-level attack.** Top-1 classification accuracy results on EPIC-Kitchens dataset under clean data and single-source adversarial perturbations on each modality, where the adversarial perturbations are generated at the feature-level, rather than the input level, as proposed in [15]. Higher is better.

Fusion	**Clean**			**Visual Perturbation**			**Motion Perturbation**			**Audio Perturbation**		
	Verb	**Noun**	**Action**	**Verb**	**Noun**	**Action**	**Verb**	**Noun**	**Action**	**Verb**	**Noun**	**Action**
Oracle (Upper Bound)	-	-	-	55.8	31.4	21.9	50.0	37.2	23.8	53.9	39.2	25.6
Concat Fusion	59.0	42.1	30.2	15.9	5.5	2.0	18.8	8.4	2.6	15.9	8.1	2.6
Mean Fusion	56.8	40.4	27.6	23.9	12.4	5.3	25.3	14.3	5.3	22.6	13.4	4.0
LEL+Robust [6]	**62.3**	42.3	31.0	31.6	23.1	6.7	38.5	32.2	14.8	31.5	28.7	9.6
Gating+Robust [5, 4]	61.9	**44.7**	**31.9**	35.7	25.9	9.3	45.9	36.2	19.5	38.2	28.8	11.8
Ours	60.8	43.5	31.3	**45.1**	**32.1**	**17.5**	**51.5**	**38.0**	**24.5**	**56.6**	**40.0**	**27.7**
Δ-Clean	**1.8**	**1.4**	**1.1**	**21.2**	**19.7**	**12.2**	**26.2**	**23.7**	**19.2**	**34.0**	**26.6**	**23.7**
Δ-Robust	-1.5	-1.2	-0.6	**9.4**	**6.2**	**8.2**	**5.6**	**1.8**	**5.0**	**18.4**	**11.2**	**15.9**

Table 6. **Untargeted attack.** Top-1 classification accuracy results on EPIC-Kitchens dataset under clean data and single-source adversarial perturbations on each modality, where the adversarial perturbations are generated to target the verb and noun class of a randomly sampled clip from the dataset, rather than the real verb and noun class. Higher is better.

C. Other Types of Attacks

In the main text of the paper, we showed across three benchmark tasks that standard multimodal models are not robust to single-source adversaries (*e.g.*, end-to-end, untargeted white box attacks on the entire multimodal model) and subsequently proposed a defense strategy based on odd-one-out learning for robust feature fusion. Here, we show on the EPIC-Kitchens dataset that our results also hold for other types of attacks: transfer attacks, targeted attacks, and feature-level attacks [15].

Transfer attacks. We consider unimodal transfer attacks on the multimodal models by (1) training a unimodal classifier on top of pretrained feature extractors for each modality, (2) generating a white-box untargeted attack from this unimodal model, (3) applying the perturbed input to the multimodal model along with the two unperturbed inputs from the other modalities. Note that the perturbed input is generated without knowledge of the fusion model and without knowledge of the unperturbed inputs from other modalities. From the performance results of different models in Supplementary Table 4, we draw two conclusions. First, standard multimodal models based on concatenation fusion ("Concat Fusion") or mean fusion ("Mean Fusion") are also not robust to these transfer attacks, which are weaker than end-to-end white box attacks; in fact, the performance is close to zero, which is no better than a unimodal model defending against the same attack. In other words, an adversary can successfully attack a single modality of a multimodal model without knowledge of how features are fused between different modalities and without knowledge of the unperturbed inputs from other modalities. Second, our approach outperforms the standard models and existing state-of-the-art robust methods on these types of attacks (see "Δ-Clean", "Δ-Robust"). We note, however, that all robust methods perform reasonably well against transfer attacks, as they are not adaptive attacks.

Feature-level attacks. Feature level attacks, as proposed in [15], are a good alternative way to assess the robustness of the fusion model, since here the attacker does not use the vulnerability of the feature extractors to generate the attack. To perform a feature-level attack, we perform a white-box attack on the feature representation for a particular modality after it has been extracted by a feature extractor g_i and before it has been fused using h (or $\tilde{h}$). We consider feature-level attacks with a maximum perturbation strength of $\epsilon = 3$ for each modality. The performance results of different models are shown in Supplementary Table 5. We note that standard multimodal models are also vulnerable to feature-level attacks; in fact, the performance is close to zero, which is no better than a unimodal model defending against the same attack. Additionally, our approach significantly outperforms the standard models and existing state-of-the-art robust methods on these types of attacks (see "Δ-Clean", "Δ-Robust").

Targeted attacks. To perform a targeted attack, we first randomly select a verb and noun class from the dataset and denote this target label as $\hat{y}$; we denote the predicted verb and noun for the input $\mathbf{x}$ as y. Subsequently, we generate a white-box adversarial attack against the model by optimizing Equation (1) of the main text, taking

$$\mathcal{L}(f(x_i + \delta, \mathbf{x}_{-i}), \hat{y}) = -\log \mathbb{P}_f(y = \hat{y}; x_i + \delta, \mathbf{x}_{-i}).$$

The performance results of different models are shown in Supplementary Table 6. Importantly, our approach also outperforms the standard models and existing state-of-the-art robust methods on these types of attacks (see "Δ-Clean", "Δ-Robust"). Note that the standard models appear to perform better than zero here; however, this is mainly because there is a chance of sampling targeted verb and noun classes that are the same as the real labels.

D. Clean performance of robust fusion strategy on KITTI

Features that are trained on perturbed data are known to perform *significantly* worse on unperturbed data [7]. In contrast, our approach is built on top of feature extractors pretrained on clean data and does not notably degrade clean performance. In the main text, we show across all three benchmarks that our performance on clean (unperturbed) data is comparable with fusion models with standard training (see the "Clean" column of the "Δ-Clean" rows in Tables 2,3,4 of the main text). Clean performance seems slightly degraded on the KITTI dataset, but these differences are not significant when standard deviations are taken into account as shown in Supplementary Table 7.

Fusion	**Clean Performance**		
	Car	**Pedestrian**	**Cyclist**
Non-Robust Model	92.8 ± 1.3	81.0 ± 4.7	84.9 ± 4.7
Ours	91.3 ± 2.7	80.8 ± 5.0	86.2 ± 4.5

Table 7. Clean performance of our robust model v. standard multimodal model, trained for the same number of epochs. The uncertainties shown are standard deviations computed over three splits of the validation data.

	Verb	**Noun**	**Action**
Vision	39.6 / 83.1	35.4 / 60.8	18.7 / 54.2
Motion	53.2 / 83.4	28.6 / 53.5	19.0 / 48.4
Audio	40.7 / 77.3)	20.6 / 40.9	12.1 / 35.9
V+M	53.9 / 85.4	39.2 / 64.6	25.6 / 58.2
V+A	50.0 / 84.3	37.2 / 63.4	23.8 / 55.9
M+A	55.8 / 84.6	31.4 / 57.8	21.9 / 52.2
V+M+A	59.0 / 86.1	42.1 / 66.5	30.2 / 60.8

Table 8. **EPIC-Kitchens Multimodal Benchmark Task**. Top-1 / Top-5 classification accuracy of standard 1-modal, 2-modal, and 3-modal models on the Epic-Kitchens verb, noun, and action recognition task. Rows 2-4 show the performance of 1-modal models based on individual modalities; rows 5-7 show the performance of 2-modal models on every pair of modalities, and row 8 shows the performance of the 3-modal model on all of the modalities.

E. Analysis of Multimodal Benchmark Tasks

In order to leverage consistency between modalities in a multimodal task to perform robust fusion, there must be shared or redundant information between them. Here, we analyze the extent to which the modalities in EPIC-Kitchens, KITTI, and CMU-MOSI are redundant, by considering the clean performance of 1-modal (unimodal), 2-modal, and 3-modal models on the task. The results in this section help shed light on the discussion of the odd-one-out network performance in the main text. Specifically, it is easier for the odd-one-out network to detect the perturbed modality for datasets such as KITTI where there is large overlap in information between data modalities; on the other hand, odd-one-out learning is more challenging on datasets such as EPIC-Kitchens where there is a fair amount of complementary information between data modalities.

EPIC-Kitchens Action Recognition. The results of our assessment of the EPIC-Kitchens multimodal benchmark are shown in Supplementary Table 8. We observe a trend of diminishing returns on clean performance as we add more modalities to the model, which indicates that there is redundant information between the modalities— notice that the drop of performance is relatively small from the 3-modal models (row 8) to any of the 2-modal models (rows 5-7), especially for the verb and noun classification tasks. How-

	Car	Pedestrian	Cyclist
RGB	93.3 / 92.9 / 86.0	82.3 / 76.3 / 71.9	90.3 / 83.0 / 80.4
Velodyne	89.6 / 87.0 / 81.0	82.2 / 76.2 / 72.1	91.7 / 84.6 / 81.8
Stereo Depth	93.9 / 88.3 / 81.6	79.8 / 71.7 / 67.8	83.8 / 80.4 / 75.7
R+V	93.3 / 92.8 / 86.2	85.0 / 80.5 / 76.3	95.1 / 87.4 / 85.1
R+S	96.1 / 93.2 / 86.5	84.6 / 79.3 / 75.0	90.3 / 85.3 / 82.7
V+S	93.5 / 90.4 / 83.8	87.0 / 80.1 / 76.0	93.4 / 86.4 / 84.0
R+V+S	96.1 / 93.5 / 86.8	85.8 / 81.5 / 77.0	93.2 / 87.7 / 83.2

Table 9. **Kitti Multimodal Benchmark Task**. Average Precision of standard 1-modal, 2-modal, and 3-modal models on Easy / Medium / Hard object detection as defined by the official KITTI evaluation server [2]. Rows 2-4 show the performance of 1-modal models based on individual modalities; rows 5-7 show the performance of 2-modal models on every pair of modalities, and row 8 shows the performance of the 3-modal model on all of the modalities.

	Two-class	Seven class
Video	69.09	39.32
Audio	59.32	33.01
Text	75.23	46.64
V+A	69.82	40.28
V+T	78.64	49.10
A+T	73.36	47.84
V+A+T	79.82	49.69

Table 10. **CMU-MOSI sentiment analysis Benchmark Task**. 2-class/7-class sentiment classification accuracy (%) of standard 1-modal, 2-modal, and 3-modal models. Rows 2-4 show the performance of 1-modal models based on individual modalities; rows 5-7 show the performance of 2-modal models on every pair of modalities, and row 8 shows the performance of the 3-modal model on all of the modalities.

ever, there is also a notable amount of complementary information between the modalities. For example, the visual modality contains more information for noun classification: the 1-modal vision model outperforms the 1-modal motion and audio models in noun accuracy, and the drop from the 3-modal model to the 2-modal model excluding vision yields the largest drop in noun accuracy. Likewise, the motion modality contains more information for verb classification: the 1-modal motion model outperforms the 1-modal vision and audio models in verb accuracy, and the drop from the 3-modal model to the 2-modal model excluding motion yields the largest drop in verb accuracy. Our findings are consistent with the observations in [3].

KITTI 2D Object Detection. The results of our assessment of the KITTI multimodal benchmark are shown in Supplementary Table 9. Here, we observe that each modality achieves high performance on the task individually. The improvement of clean performance when increasing the number of modalities from 1 (rows 2-4) to 3 (row 8) is rather minimal. This demonstrates that the information between these three modalities is highly redundant, and there is little complementary information between the modalities.

CMU-MOSI Sentiment Analysis. The results of our assessment of CMU-MOSI benchmark are shown in Supplementary Table 10. In this task, text is the modality that carries most information and adding video improves the performance more significantly than adding audio. Comparing row 8 which uses all three modalities and row 7 which uses video and text, as well as comparing row 4 which uses video and audio with row 1 which uses only video, one can see that there exist strong redundancy between audio and video.

F. Early Fusion vs. Late Fusion Models

In our preliminary experiments, we found that early fusion models are notably less robust than late fusion methods, which motivated us to focus on late fusion approaches in this work. Supplementary Table 11 shows our results comparing early and late fusion approaches on action recognition on the UCF-101 dataset. Here the input consists of 2-5 frames (which are treated as different views) and only one of the frames is adversarially perturbed while the others are clean An early fusion method that concatenates the frames over their spatial dimensions before passing them through a convolutional neural network is significantly less robust than a late fusion network that processes the frames individually and averages the output.

We also evaluated the robustness of transformers that perform early fusion of video, audio, and text from the MOSI dataset, and found that they performed notably worse than late fusion models as shown in Supplementary Table 12.

One intuitive explanation for these results is the following: in early fusion models, there are multiple early interactions between attacked and clean modalities, so feature extraction becomes strongly influenced by the attacked input. In contrast, in late fusion models, feature extraction of the clean modalities is mostly shielded from the attack.

G. Additional Tables

In the following, we provide additional results that were deferred from the main text due to space constraints. Sup-

		Number of Input Frames			
	Input Type	**2**	**3**	**4**	**5**
Early Fusion Model	Clean frames	69.7/89.3	70.2/89.1	68.3/87.2	70.9/87.4
	Single-frame attack	16.3/38.1	20.8/45.1	20.0/45.0	22.5/48.8
Late Fusion Model	Clean frames	72.4/89.0	73.0/90.0	74.0/89.6	73.5/90.5
	Single-frame attack	46.9/76.4	62.0/83.5	66.7/85.8	67.0/87.4

Table 11. **Early fusion vs. late fusion robustness on UCF-101 action recognition.** The early fusion model is significantly less robust than the late fusion model under a single-frame perturbation. Top 1 / Top 5 classification accuracy is shown (higher is better). See text for details.

	Audio	**Video**	**Text**
Early Fusion (transformer) Model	15.9	18.2	27.5
Late Fusion Model	56.92	51.23	39.1

Table 12. **Early fusion vs. late fusion robustness on two-class MOSI sentiment analysis.** The early fusion model is significantly less robust than the late fusion model, especially under audio and video perturbation.

plementary Table 13 provides top-1 and top-5 accuracy of multimodal models on action recognition on EPIC-Kitchens and supplement the result of Table 2 in the main text. Supplementary Table 14 shows average precision performance of multimodal models on easy/medium/hard object detection on the KITTI dataset and supplement the results of Table 3 in the main text. Supplementary Table 15 shows end-to-end adversarial training results on MOSI – the proposed method outperforms adversarial training on clean input for over 20% with small sacrifice on robust accuracy.

References

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018. 1

[2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1, 5

[3] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5492–5501, 2019. 1, 5

[4] Jaekyum Kim, Jaehyung Choi, Yechol Kim, Junho Koh, Chung Choo Chung, and Jun Won Choi. Robust camera lidar sensor fusion via deep gated information fusion network. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1620–1625. IEEE, 2018. 3, 7

[5] Jaekyum Kim, Junho Koh, Yecheol Kim, Jaehyung Choi, Youngbae Hwang, and Jun Won Choi. Robust deep multimodal learning based on gated information fusion network. In *Asian Conference on Computer Vision*, pages 90–106. Springer, 2018. 3, 7

[6] Taewan Kim and Joydeep Ghosh. On single source robustness in deep fusion models. In *Advances in Neural Information Processing Systems*, pages 4814–4825, 2019. 3, 7

[7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 4

[8] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 1

[9] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 1

[10] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1, 2

[11] Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1085–1097, 2019. 2

[12] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 1

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[14] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 1

[15] Qiuling Xu, Guanhong Tao, Siyuan Cheng, Lin Tan, and Xiangyu Zhang. Towards feature space adversarial attack. *arXiv preprint arXiv:2004.12385*, 2020. 3, 4

Fusion	Clean			Visual Perturbation			Motion Perturbation			Audio Perturbation		
	Verb	**Noun**	**Action**	**Verb**	**Noun**	**Action**	**Verb**	**Noun**	**Action**	**Verb**	**Noun**	**Action**
Concat Fusion	59.0 (86.1)	42.1 (66.5)	30.2 (60.8)	0.1 (25.0)	0.0 (7.3)	0.0 (4.0)	0.2 (24.6)	0.0 (5.1)	0.0 (2.7)	0.1 (21.2)	0.0 (6.8)	0.0 (3.5)
Mean Fusion	56.8 (85.9)	40.4 (66.1)	27.6 (59.7)	0.3 (58.1)	0.8 (15.8)	0.0 (11.8)	0.3 (64.2)	0.3 (14.3)	0.0 (12.0)	0.4 (48.0)	0.3 (16.9)	0.0 (12.0)
LEL+Robust [6]	61.2 (86.7)	**43.1 (68.9)**	30.5 (62.3)	22.3 (76.3)	11.6 (37.2)	6.6 (34.9)	25.4 (78.6)	24.6 (52.3)	12.0 (46.5)	20.4 (77.6)	17.7 (47.5)	8.0 (43.2)
Gating+Robust [5, 4]	60.9 (87.6)	43.0 (68.7)	30.6 (62.8)	26.0 (77.9)	10.9 (42.0)	6.2 (39.0)	35.9 (83.6)	26.9 (58.3)	14.3 (52.5)	21.3 (79.9)	16.2 (51.4)	7.0 (47.2)
Ours	**61.5 (88.2)**	42.5 (68.4)	**31.4 (63.0)**	**48.0 (83.2)**	**24.2 (53.2)**	**16.8 (48.9)**	**48.5 (85.2)**	**35.6 (63.8)**	**22.1 (57.4)**	**46.5 (85.2)**	**33.3 (62.3)**	**22.1 (57.1)**

Table 13. Top-1 (Top-5) classification accuracy results on EPIC-Kitchens dataset under clean data and single-source adversarial perturbations on each modality. Higher is better. See Table 2 in main paper and accompanying details in text for details.

Fusion	**Clean**			**Visual (RGB) Perturbation**		
	Car	**Ped**	**Cyc**	**Car**	**Ped**	**Cyc**
Concat Fusion	96.1 / 93.5 / 86.8	85.7 / 81.5 / 77.0	93.2 / 87.7 / 83.2	15.6 / 14.3 / 14.8	13.6 / 10.7 / 10.3	13.8 / 12.3 / 12.9
Mean Fusion	96.5 / 93.6 / 86.6	84.3 / 77.7 / 73.4	91.9 / 86.7 / 81.8	13.2 / 12.6 / 13.1	18.1 / 15.2 / 14.2	11.9 / 10.5 / 10.2
LEL+Robust [6]	80.5 / 71.4 / 67.3	69.0 / 64.2 / 61.3	75.7 / 80.0 / 75.3	3.71 / 3.95 / 4.62	16.9 / 15.4 / 14.3	16.4 / 13.9 / 13.2
Gating+Robust [5, 4]	90.6 / 89.4 / 82.8	81.5 / 74.7 / 72.6	92.9 / 84.6 / 81.8	67.3 / 57.2 / 53.1	62.0 / 54.2 / 50.7	68.6 / 56.0 / 53.1
Ours	95.6 / 90.6 / 83.9	84.5 / 79.9 / 75.7	90.4 / 85.4 / 80.6	**89.6 / 85.1 / 78.9**	**80.5 / 73.9 / 69.8**	**87.9 / 82.3 / 77.6**

Fusion	**Depth (Velo) Perturbation**			**Stereo Disparity Perturbation**		
	Car	**Ped**	**Cyc**	**Car**	**Ped**	**Cyc**
Concat Fusion	3.43 / 1.58 / 1.59	11.3 / 11.1 / 11.4	8.72 / 8.82 / 8.22	7.37 / 3.57 / 3.44	8.08 / 4.64 / 4.36	9.13 / 7.23 / 7.72
Mean Fusion	6.77 / 3.16 / 2.90	13.7 / 12.9 / 12.8	10.1 / 7.88 / 8.07	6.88 / 3.08 / 2.73	9.17 / 8.03 / 8.81	12.2 / 7.77 / 7.28
LEL+Robust [6]	7.30 / 6.83 / 6.73	24.4 / 20.6 / 18.9	28.8 / 24.8 / 24.7	10.1 / 9.39 / 9.50	26.0 / 24.2 / 21.8	25.7 / 24.7 / 23.8
Gating+Robust [5, 4]	51.7 / 46.5 / 43.2	53.5 / 45.7 / 42.1	58.8 / 45.6 / 53.8	43.9 / 41.6 / 38.8	53.9 / 47.4 / 44.1	60.0 / 48.8 / 46.8
Ours	**92.8 / 87.8 / 79.4**	**78.3 / 71.1 / 67.1**	**88.9 / 85.8 / 81.1**	**92.8 / 89.8 / 83.1**	**83.6 / 76.8 / 72.4**	**88.1 / 84.7 / 79.9**

Table 14. Evaluation of Average Precision for 2D object detection (easy/medium/hard difficulty) on the KITTI dataset under clean data and single-source adversarial perturbations on each modality. Higher is better. See Table 3 in main paper and accompanying details in text for details.

Fusion	**Clean**		**Audio Perturbation**		**Video Perturbation**		**Text Perturbation**	
	2-class	**7-class**	**2-class**	**7-class**	**2-class**	**7-class**	**2-class**	**7-class**
Adversarial training	61.23	36.02	74.38	42.13	70.34	39.28	69.94	32.45
Ours	**82.03**	**50.89**	73.18	42.06	69.94	38.20	66.13	30.20

Table 15. Binary and seven-class classification results (%) of end-to-end adversarial training and our method on MOSI. Higher is better.