

APPENDIX A

MORE IMPLEMENTATION DETAILS IN SECTION 1

We provide the implementation details in this part for the experiment in Section 1. This experiment is conducted on CREMA-D dataset. The comparison methods include vanilla MML, HA-based MML, and SA-based MML. Suppose that the input data is $(\mathbf{x}^a, \mathbf{x}^v)$, given an audio encoder $\phi^a(\cdot)$ and a video encoder $\phi^v(\cdot)$, the features are calculated by:

$$\forall o \in \{a, v\}, \mathbf{z}^o = \phi^o(\mathbf{x}^o; \Phi^o). \quad (1)$$

Then, we utilize the concatenation function for fusion and a one-layer fully connected network for classification. In other words, our classifier is conducted by a linear projection.

After taking the softmax layer, the prediction can be calculated by:

$$\forall o \in \{a, v\}, \hat{\mathbf{h}}^o = \mathbf{W}^o \mathbf{z}^o + \frac{\mathbf{b}}{2}, \quad (2)$$

$$\hat{\mathbf{y}}^o = \text{softmax}(\hat{\mathbf{h}}^o), \quad (3)$$

$$\hat{\mathbf{y}} = \text{softmax}(\hat{\mathbf{h}}^a + \hat{\mathbf{h}}^v), \quad (4)$$

where $\hat{\mathbf{y}}, \hat{\mathbf{y}}^a, \hat{\mathbf{y}}^v$ denote the total prediction, implicit audio prediction and implicit video prediction, respectively.

For vanilla MML, the objective function can be written as:

$$L_{\text{vanilla}}(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \mathbf{y}^\top \log(\hat{\mathbf{y}}). \quad (5)$$

As the optimizing of Problem 5, the implicit audio prediction $\hat{\mathbf{y}}^a$ and implicit video prediction $\hat{\mathbf{y}}^v$ are aligned to ground-truth $\mathbf{y}$ and further expected to well-aligned with each other.

For HA-based MML, we further integrate two unimodal losses to improve the performance. Concretely, we utilize two classifiers for unimodal prediction. The predictions are defined as:

$$\begin{aligned} \forall o \in \{a, v\}, \tilde{\mathbf{h}}^o &= \mathbf{W}^o \mathbf{z}^o + \mathbf{b}^o, \\ \tilde{\mathbf{y}}^o &= \text{softmax}(\tilde{\mathbf{h}}^o). \end{aligned} \quad (6)$$

Based on Eq. (6), the objective function of HA-based MML is defined as:

$$L_{\text{HA}}(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \mathbf{y}^\top \log(\hat{\mathbf{y}}) + \underbrace{\frac{1}{|\mathcal{D}|} \sum_{o \in \{a, v\}} \sum_{\mathbf{x}^o \in \mathcal{D}} \mathbf{y}^\top \log(\tilde{\mathbf{y}}^o)}_{\text{Implicit Modality Alignment}}. \quad (7)$$

In Eq. (7), the unimodal prediction $\tilde{\mathbf{y}}^a$ and $\tilde{\mathbf{y}}^v$ are required to align to $\mathbf{y}$. Ideally, they are expected to well-aligned to each other. We denote this procedure as:

$$\text{Hard Alignment: } \hat{\mathbf{y}}^a, \hat{\mathbf{y}}^v \rightsquigarrow \mathbf{y} \Rightarrow \hat{\mathbf{y}}^a \rightsquigarrow \hat{\mathbf{y}}^v,$$

Then, for SA-based MML, the objective function is defined as:

$$L_{\text{SA}}(\mathcal{D}) = \underbrace{\frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \hat{\mathbf{y}}^\top \log(\mathbf{y})}_{\text{Modality Fusion}} + \underbrace{\frac{1}{2} \mathcal{D}_{\text{KL}}(\hat{\mathbf{y}}^a \parallel \bar{\mathbf{y}}) + \frac{1}{2} \mathcal{D}_{\text{KL}}(\hat{\mathbf{y}}^v \parallel \bar{\mathbf{y}})}_{\text{Explicit Modality Alignment}}, \quad (8)$$

where $\bar{\mathbf{y}} \triangleq \frac{1}{2}(\mathbf{y}^a + \mathbf{y}^v)$. In Eq. (8), the unimodal predictions are expected to align with each other with soft label $\bar{\mathbf{y}}$. We define this procedure as:

$$\text{Soft Alignment: } \hat{\mathbf{y}}^a, \hat{\mathbf{y}}^v \rightsquigarrow \bar{\mathbf{y}} \Rightarrow \hat{\mathbf{y}}^a \rightsquigarrow \hat{\mathbf{y}}^v,$$

We provide the detailed architecture for vanilla MML, HA-based MML, and SA-based MML in Figure A.

APPENDIX B

PREDICTION LEVEL ALIGNMENT WITH ARCHITECTURE CONTAINING UNIMODAL LOSSES

A more general architecture widely adopted by existing works [1, 2] is a multimodal loss with unimodal losses. We show that how to incorporate modality alignment with modality fusion for these methods in this section. The architecture is shown in Figure B. For this architecture, the individual classifiers are used to obtain unimodal predictions $\bar{\mathbf{y}}^a$ and $\bar{\mathbf{y}}^v$. The modality alignment loss can be applied as follows:

$$L_{\text{Align}} = L_{\text{JS}}(\bar{\mathbf{y}}^a, \bar{\mathbf{y}}^v).$$

We conduct an experiment to compare the alignment strategy for this architecture. The results in Table A show that the method with this architecture can achieve better performance compared the method with one layer network based classifier. For fair comparison with existing approach, we adopt the method with one layer network in this paper.

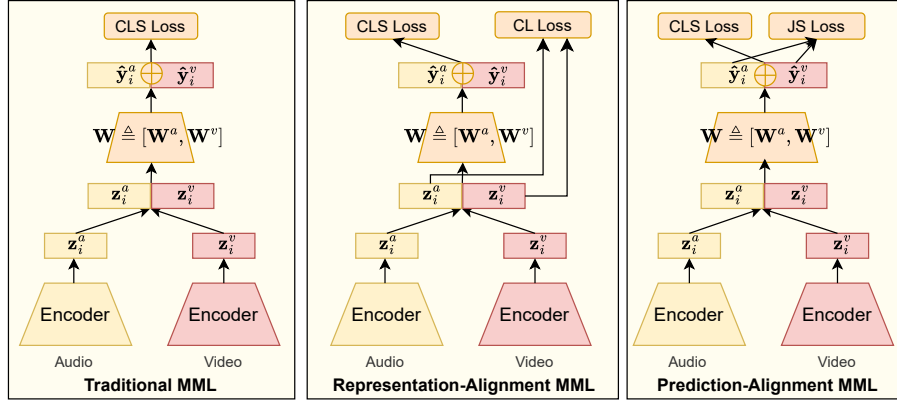


Fig. A. Illustration of architecture for vanilla MML, HA-based MML, and SA-based MML.

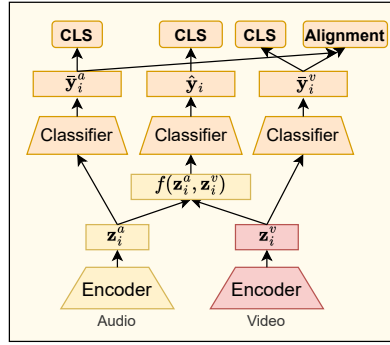


Fig. B. Modality alignment for architecture containing unimodal losses.

APPENDIX C

ADDITIONAL EXPERIMENTAL DETAILS AND RESULTS

C.1 Datasets and Implementation Details

Dataset Description and Splitting: Among these datasets, the KSounds, CREMA-D and VGGSound datasets consist of both audio and video modalities. The KSounds dataset, which contains 19,000 video clips categorized into 31 distinct actions, aims to advance video action recognition. The CREMA-D dataset, encompassing 7,442 clips, is divided into six emotional categories to enhance speech emotion analysis. The VGGSound dataset, which contains 310 classes and a wide range of audio events in everyday life, is a relatively large dataset. Furthermore, the Sarcasm and Twitter2015 datasets consist of image and text modalities. The Sarcasm dataset offers a compilation of 24,635 text-image pairs. The Twitter2015 dataset contains 5,338 text-image combinations from Twitter. Lastly, NVGesture and IEMOCAP datasets are used to construct research that goes beyond the limitation of two modalities. In this paper, we use RGB, Depth, as well as optical flow (OF) modalities for NVGesture and audio, video, as well as text for IEMOCAP datasets.

For the datasets with audio and video modalities, the KSounds dataset is divided into a training set of 15,000 clips, a validation set of 1,900 clips, and a test set of 1,900 clips. CREMA-D dataset is split into a training set with 6,698 clips and a testing set with 744 clips. VGGSound dataset is divided into a training set with 168,618 videos, a validation set with 33,732 videos, and a testing set with 13,954 videos. For datasets with image and text modalities, the Sarcasm dataset is divided into a training set with 19,816 pairs, a validation set with 2,410 pairs, and a testing set with 2,409 pairs. Twitter2015 dataset is divided into a training set with 3,179 pairs, a validation set with 1,122 pairs, and a testing set with 1,037 pairs. For NVGesture dataset, we divide it into a training set with 1,050 data points and a testing set with 482 data points. The IEMOCAP dataset is partitioned into 3838 dialogues for training, 838 dialogues for validation, and 1098 dialogues for testing.

Implementation Details: In our experiments, we utilize raw data as inputs. Following [1, 3], for the KSounds and CREMA-D datasets, ResNet18 serves as the foundational architecture for processing both audio and video data. For video analysis, we select 10 frames from each clip and subsequently sample three frames uniformly as inputs. We adapt ResNet18’s input channels from three to one to accommodate our data format [4]. In terms of audio, we transform our sound recordings into spectrograms measuring 257×1004 for KSounds and 257×299 for CREMA-D, employing the librosa [5] library for conversion. For text-image datasets, our framework incorporates ResNet50 for images and BERT [6] for text processing.

TABLE A
Comparison for the methods with different architectures.

Architecture	Audio	Video	Multiple
w/ alignment	66.94%	65.19%	81.59%
w/o alignment	68.55%	64.38%	77.96%

TABLE B
Ablation study on datasets with three datasets. The symbols “PA” and “RA” denote whether the prediction alignment, representation alignment, and dynamic integration are applied during training.

Dataset	Module			MAP/F1			
	PA	RA	DI	Multiple	Unimodal1	Unimodal2	Unimodal3
NVGesture	×	×	×	81.98%	6.24%	76.79%	6.54%
	✓	×	×	83.86%	70.60%	76.82%	78.85%
	✓	×	✓	84.85%	73.24%	80.08%	79.88%
	×	✓	×	83.78%	72.66%	77.35%	75.30%
	×	✓	✓	84.68%	75.52%	82.91%	80.31%
	✓	✓	×	84.93%	73.05%	80.22%	80.00%
	✓	✓	✓	85.65%	78.42%	83.70%	81.21%
IEMOCAP	×	×	×	84.09%	24.64%	66.69%	56.42%
	✓	×	×	85.37%	26.54%	74.58%	61.74%
	✓	×	✓	85.76%	24.86%	75.72%	63.85%
	×	✓	×	85.74%	25.96%	76.72%	61.78%
	×	✓	✓	85.99%	26.64%	77.07%	61.79%
	✓	✓	×	85.82%	26.69%	76.52%	62.12%
	✓	✓	✓	86.04%	26.80%	77.03%	63.68%

We resize images to 224×224 and limit text sequences to a maximum length of 128 characters. Optimization for the audio-video datasets is conducted using stochastic gradient descent (SGD) with a momentum set to 0.9 and a weight decay parameter of 10^{-1} . We initialize the learning rate to 10^{-2} , progressively reducing it by a factor of ten upon observing a plateau in loss reduction, with a batch size of 256. For text-image datasets [7, 8], we employ the Adam optimizer starting with a learning rate of 10^{-4} , with a batch size of 128. For our methods, we utilize “Ours-H” and “Ours-LB” to denote the heuristic strategy based approach, dynamic learning-based approach with hybrid alignment. All models are trained on a single RTX 3090 GPU.

C.2 Ablation Study on the Datasets with Three Modalities

To further demonstrate the effectiveness of the key components of our method, i.e., the prediction-level alignment, representation-level alignment, and dynamic integration, we conduct ablation study experiments on NVGesture and IEMOCAP datasets. The results are reported in Table B, where the symbols are defined analogously in Table 4.

From the results in Table B, we can find that: (1). Both prediction-level alignment and representation-level alignment can boost performance compared MML method w/o PA, RA, and DI; (2). By integrating prediction-level alignment and representation-level alignment simultaneously, the performance can be further improved; (3). By dynamically balancing modal fusion and alignment, the dynamic integration module effectively enhances the overall performance of multimodal learning.

TABLE C
Comparison under semi-supervised scenario on VGGSound dataset.

Method	$\mu = 10\%$		$\mu = 20\%$		$\mu = 30\%$		$\mu = 40\%$	
	Audio Acc	Video MAP	Audio Acc	Video MAP	Audio Acc	Video MAP	Audio Acc	Video MAP
Vanilla SML	31.24%	32.51%	28.88%	29.87%	30.62%	33.07%	34.43%	36.58%
MLA w/ SSL	<u>34.75%</u>	35.59%	33.08%	34.71%	<u>35.40%</u>	<u>37.48%</u>	39.14%	40.77%
MMPareto w/ SSL	32.59%	34.32%	31.30%	33.83%	32.46%	36.08%	36.11%	37.39%
ReconBoost w/ SSL	33.01%	<u>36.90%</u>	32.66%	34.62%	33.20%	36.18%	37.46%	39.79%
DSRL	33.72%	35.01%	31.24%	31.27%	34.14%	36.16%	39.08%	40.78%
SMDDRL	33.60%	35.40%	32.17%	34.28%	35.46%	36.91%	<u>39.43%</u>	41.04%
CorMulT	32.49%	34.83%	32.78%	35.07%	34.91%	37.35%	38.53%	40.50%
DTC	33.95%	35.89%	33.37%	<u>35.19%</u>	35.17%	36.88%	38.92%	40.72%
Ours-LB	35.01%	37.48%	<u>33.20%</u>	36.07%	35.30%	37.50%	39.53%	<u>41.02%</u>

C.3 Semi-Supervised Learning Comparison with Different Sampling Proportion μ

To further study the effectiveness of our method under semi-supervised learning scenario, we report the performance comparison for our method and competitive baselines with different μ on VGGSound dataset in Table C. The results demonstrate that our method can achieve the best performance in almost all cases with different μ .

APPENDIX D

THEORETICAL ANALYSIS

D.1 Notations

For clarity of exposition, we first introduce the notation used throughout this section. We denote $\langle \cdot, \cdot \rangle$ as inner product, i.e., $\langle \mathbf{x}, \mathbf{y} \rangle \triangleq \mathbf{x}^T \mathbf{y}$. The Euclidean norm of vector $\mathbf{x} \in \mathcal{R}^d$ is denoted as: $\|\mathbf{x}\| \triangleq \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. And the matrix norm of $\mathbf{A} \in \mathcal{R}^{m \times d}$ refers to the Frobenius norm of $\mathbf{A}$, i.e., $\|\mathbf{A}\| \triangleq \max\{\|\mathbf{A}\mathbf{x}\| : \mathbf{x} \in \mathcal{R}^d, \|\mathbf{x}\| \leq 1\}$, which is equivalent to the maximum singular value of $\mathbf{A}$. The expectation of a random variable $\mathbf{X}$ is denoted as $\mathbb{E}[\mathbf{X}]$. Given two symmetric matrix $\mathbf{A}, \mathbf{B} \in \mathcal{R}^{n \times n}$, $\mathbf{A} \preceq \mathbf{B} \Leftrightarrow \mathbf{B} - \mathbf{A}$ is a non negative definite matrix. If a function $f \in C^2[\mathcal{S}]$, that means f is twice continuously differentiable on $\mathcal{S}$. We denote $\zeta_{[t]}$ as the set $\{\zeta_0, \zeta_1, \dots, \zeta_{t-1}\}$. We denote the stochastic vector obtained by accessing the stochastic gradient oracle of a function f as $\tilde{\nabla} G_f(\mathbf{x}, \zeta)$, which is used to approximate $\nabla f(\mathbf{x})$ at each iteration of the algorithm, where ζ is a random variable and is usually required to be historically independent. Furthermore, the definitions of the notations for partial derivatives required in our analysis are as follows:

$$\nabla_{\Theta} L_{\text{Fusion}}(\Theta) \triangleq \left[\frac{\partial L_{\text{Fusion}}}{\partial \Theta_1}, \dots, \frac{\partial L_{\text{Fusion}}}{\partial \Theta_d} \right]^T \quad (9)$$

$$\nabla_{\Theta}^2 L_{\text{Fusion}}(\Theta) \triangleq \text{Hessian matrix of } L_{\text{Fusion}} \text{ w.r.t } \Theta \quad (10)$$

$$\nabla_{\alpha\Theta}^2 L_{\text{Total}}(\Theta) \triangleq \nabla_{\Theta} \left(\frac{\partial L_{\text{Total}}}{\partial \alpha} \right) = \nabla_{\Theta} L_{\text{Align}}(\Theta) - \nabla_{\Theta} L_{\text{Fusion}}(\Theta). \quad (11)$$

$$\nabla_{\Theta}^2 L_{\text{Total}}(\alpha, \Theta) \triangleq (1 - \alpha) \nabla_{\Theta}^2 L_{\text{Fusion}}(\Theta) + \alpha \nabla_{\Theta}^2 L_{\text{Align}}(\Theta) \quad (12)$$

$$\nabla L_{\text{Fusion}}(\alpha, \Theta^*(\alpha)) \triangleq \frac{d L_{\text{Fusion}}(\Theta^*(\alpha))}{d\alpha} = -\nabla_{\alpha\Theta}^2 L_{\text{Total}}(\Theta^*(\alpha)) [\nabla_{\Theta}^2 L_{\text{Total}}(\alpha, \Theta^*(\alpha))]^{-1} \nabla_{\Theta} L_{\text{Fusion}}(\Theta^*(\alpha)) \quad (13)$$

$$\nabla L_{\text{Fusion}}(\alpha, \Theta) \triangleq -\nabla_{\alpha\Theta}^2 L_{\text{Total}}(\Theta) [\nabla_{\Theta}^2 L_{\text{Total}}(\alpha, \Theta)]^{-1} \nabla_{\Theta} L_{\text{Fusion}}(\Theta) \quad (14)$$

D.2 Convergence Analysis

We now present definitions of several key concepts employed in our theoretical analysis for the Algo. 1.

We first define the Lipschitz continuous.

Definition D.1. (Lipschitz Continuous) A function $f : \mathcal{R}^{d_1} \mapsto \mathcal{R}^{d_2 \times d_3}$ is L -Lipschitz continuous ($L > 0$) on $\mathcal{S}$, if

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{S} \quad (15)$$

where $\|\cdot\|$ denotes norm of a vector or matrix we defined depending on whether d_3 equal to 1.

Then we have:

Proposition D.1. If a function $f : \mathcal{R}^d \mapsto \mathcal{R}$ is a L -smooth function on $\mathcal{S}$, then $\forall \mathbf{x}, \mathbf{y} \in \mathcal{S}$, we have

$$f(\mathbf{y}) - f(\mathbf{x}) \leq \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2 \quad (16)$$

Then we define the strongly convex.

Definition D.2. (μ -Strongly Convex) A function $g : \mathcal{R}^d \mapsto \mathcal{R}$ is μ -strongly convex ($\mu > 0$) on $\mathcal{S}$, if $\forall \mathbf{x}, \mathbf{y} \in \mathcal{S}$ and $\lambda \in [0, 1]$

$$\lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}) \geq f(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}) + \frac{\mu}{2} \lambda (1 - \lambda) \|\mathbf{y} - \mathbf{x}\|^2, \quad (17)$$

If f is differentiable, inequality (17) is equivalent to $f(\mathbf{x}) - f(\mathbf{y}) \geq \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|^2$. If $f \in C^2[\mathcal{S}]$ and μ -strongly convex, we can get $\mu \mathbf{I} \preceq \nabla^2 f(\mathbf{x})$.

Then, we define the unbiased stochastic gradient oracle and variance-bounded stochastic gradient oracle.

Definition D.3. (Unbiased Stochastic Gradient Oracle) A stochastic gradient oracle of a function f is unbiased, if

$$\mathbb{E}_{\zeta} [\tilde{\nabla} G_f(\mathbf{x}, \zeta)] = \nabla f(\mathbf{x}) \quad (18)$$

Definition D.4. (Variance-Bounded Stochastic Gradient Oracle) A stochastic gradient oracle is variance-bounded, if $\exists \sigma > 0$

$$\mathbb{E}_{\zeta} [\|\tilde{\nabla} G_f(\mathbf{x}, \zeta) - \nabla f(\mathbf{x})\|^2] \leq \sigma^2, \quad \forall \mathbf{x} \in \mathcal{S} \quad (19)$$

Based on the aforementioned definitions, we make the following assumptions.

Assumption D.2. (Deterministic)

- (a) $\nabla_{\Theta} L_{\text{Fusion}}(\Theta)$ is $\mathcal{L}_{\Theta}^{\text{Fusion}}$ -Lipschitz continuous w.r.t $\Theta \in \mathcal{R}^d$.
- (b) $\exists C_{\Theta}^{\text{Fusion}} > 0, \|\nabla_{\Theta} L_{\text{Fusion}}(\Theta)\| \leq C_{\Theta}^{\text{Fusion}}$.
- (c) $L_{\text{Fusion}}(\Theta), L_{\text{Align}}(\Theta) \in C^2[\mathcal{R}^d]$, i.e., $L_{\text{Total}}(\alpha, \Theta) \in C^2[[0, 1] \times \mathcal{R}^d]$.

- (d) $\forall \alpha \in [0, 1], (1 - \alpha)L_{\text{Fusion}}(\Theta) + \alpha L_{\text{Align}}(\Theta)$ is strongly convex w.r.t $\Theta \in \mathcal{R}^d$. i.e., $L_{\text{Total}}(\alpha, \Theta)$ is μ_{Total} -strongly convex w.r.t $\Theta \in \mathcal{R}^d$.
- (e) $\forall \alpha \in [0, 1], L_{\text{Align}}(\Theta) - L_{\text{Fusion}}(\Theta)$ and $(1 - \alpha)\nabla_{\Theta} L_{\text{Fusion}}(\Theta) + \alpha \nabla_{\Theta} L_{\text{Align}}(\Theta)$ are both smooth functions w.r.t $\Theta \in \mathcal{R}^d$, i.e., $\nabla_{\alpha\Theta}^2 L_{\text{Total}}(\Theta)$ and $\nabla_{\Theta}^2 L_{\text{Total}}(\alpha, \Theta)$ are $\mathcal{L}_{\alpha\Theta}^{\text{Total}}$ and $\mathcal{L}_{\Theta\Theta}^{\text{Total}}$ -Lipschitz continuous w.r.t $\Theta \in \mathcal{R}^d$.
- (f) $\exists C_{\alpha\Theta}^{\text{Total}} > 0, \|\nabla_{\Theta} L_{\text{Align}}(\Theta) - \nabla_{\Theta} L_{\text{Fusion}}(\Theta)\| \leq C_{\alpha\Theta}^{\text{Total}}$. i.e., $\|\nabla_{\alpha\Theta}^2 L_{\text{Total}}(\Theta)\| \leq C_{\alpha\Theta}^{\text{Total}}$.
- (g) $\forall \Theta \in \mathcal{R}^d, \nabla_{\Theta}^2 L_{\text{Total}}(\alpha, \Theta)$ is $\tilde{\mathcal{L}}_{\Theta\Theta}^{\text{Total}}$ -Lipschitz continuous w.r.t $\alpha \in [0, 1]$.

Note that Assumption D.2 (g) is equivalent to $\|\nabla_{\Theta}^2 L_{\text{Align}}(\Theta) - \nabla_{\Theta}^2 L_{\text{Fusion}}(\Theta)\|$ is bounded.

Assumption D.3. (Stochastic)

- (a) Stochastic oracles $\tilde{\nabla}_{\Theta} G_{\text{Total}}(\alpha, \Theta, \zeta^{(1)}), \tilde{\nabla}_{\Theta} G_{\text{Fusion}}(\Theta, \xi)$ and $\tilde{\nabla}_{\alpha\Theta} G_{\text{Total}}(\alpha, \Theta, \zeta^{(2)})$ in Algo. 1 for estimate $\nabla_{\Theta} L_{\text{Total}}(\alpha, \Theta), \nabla_{\Theta} L_{\text{Fusion}}(\Theta, \xi)$ and $\nabla_{\alpha\Theta} L_{\text{Total}}(\Theta)$ are unbiased and variance bounded with $(\sigma_{\Theta}^{\text{Total}})^2, (\sigma_{\Theta}^{\text{Fusion}})^2$ and $(\sigma_{\alpha\Theta}^{\text{Total}})^2$.
- (b) Random variables $\xi, \zeta^{(1)}, \zeta^{(2)}, \zeta^{(3)}$ are independent.
- (c) The stochastic hessian inverse approximation $\mathcal{H}_{\Theta\Theta}(\alpha, \Theta, \zeta^{(3)})$ for $[\nabla_{\Theta}^2 L_{\text{Total}}(\alpha, \Theta)]^{-1}$ is unbiased and the variance bounded by $\sigma_{\mathcal{H}}^2$.

According to the assumptions, we first give a basic lemma to show that the existence and differentiability of an implicit function of $\Theta^*(\alpha)$ w.r.t $\alpha \in [0, 1]$.

Lemma D.4. If Assumption D.2 (c) and (d) hold.

- i) For any $\alpha \in [0, 1], \Theta^*(\alpha)$ is unique and continuously differentiable and we have

$$\nabla \Theta^*(\alpha) = -\mathcal{M}(\alpha, \Theta^*(\alpha))^T \quad (20)$$

- ii) $\nabla L_{\text{Fusion}}(\alpha, \Theta^*(\alpha))$ can be viewed as a function of $\alpha \in [0, 1]$, is given by

$$\nabla L_{\text{Fusion}}(\alpha, \Theta^*(\alpha)) = -\mathcal{M}(\alpha, \Theta^*(\alpha)) \nabla_{\Theta} L_{\text{Fusion}}(\Theta) \quad (21)$$

where $\mathcal{M}(\alpha, \Theta) \triangleq \nabla_{\alpha\Theta}^2 L_{\text{Total}}(\Theta) [\nabla_{\Theta}^2 L_{\text{Total}}(\alpha, \Theta)]^{-1}$.

Proof. The above results have been well-known from properties of implicit functions. For the sake of completeness, we provide a brief proof.

For i), the uniqueness of $\Theta^*(\alpha)$ can be guaranteed under the Assumption D.2 (d). $\Theta^*(\alpha)$ is continuously differentiable because the implicit function theorem (see Theorem 11.5 in [9]). Due to the definition, we have $\nabla_{\Theta} L_{\text{Total}}(\alpha, \Theta^*(\alpha)) = 0$. Then, by taking derivative on both sides, using the chain rule, and the implicit function theorem, we obtain

$$\nabla_{\alpha\Theta}^2 L_{\text{Total}}(\Theta^*(\alpha)) + \nabla_{\Theta}^2 L_{\text{Total}}(\alpha, \Theta^*(\alpha)) \nabla \Theta^*(\alpha) = 0, \quad \forall \alpha \in [0, 1]. \quad (22)$$

Just take the inverse we finish the proof of part i).

Under the Assumption D.2 (c) and the continuously differentiability of $\Theta^*(\alpha)$, L_{Fusion} is obviously continuously differentiable w.r.t α . Then by the chain rule, part ii) follows immediately. $\square$

Lemma D.5. The following statements provide powerful tools to estimate the error induced by the inner loop of Algorithm

- i) If Assumption D.2 (a)-(f) hold, then $\forall (\alpha, \Theta) \in [0, 1] \times \mathcal{R}^d$, we have

$$\|\nabla L_{\text{Fusion}}(\alpha, \Theta) - \nabla L_{\text{Fusion}}(\alpha, \Theta^*(\alpha))\| \leq D_1 \|\Theta - \Theta^*(\alpha)\| \quad (23)$$

where $D_1 \triangleq \frac{C_{\alpha\Theta}^{\text{Total}} \mathcal{L}_{\Theta}^{\text{Fusion}}}{\mu_{\text{Total}}} + \left(\frac{\mathcal{L}_{\alpha\Theta}^{\text{Total}}}{\mu_{\text{Total}}} + \frac{\mathcal{L}_{\Theta\Theta}^{\text{Total}} C_{\alpha\Theta}^{\text{Total}}}{\mu_{\text{Total}}^2} \right) C_{\Theta}^{\text{Fusion}}$

- ii) If Assumption D.2 (c), (d) and (f) hold, then $\Theta^*(\alpha)$ is $\frac{C_{\alpha\Theta}^{\text{Total}}}{\mu_{\text{Total}}}$ -Lipschitz continuous w.r.t α , i.e.,

$$\|\Theta(\alpha_1) - \Theta(\alpha_2)\| \leq \frac{C_{\alpha\Theta}^{\text{Total}}}{\mu_{\text{Total}}} \|\alpha_1 - \alpha_2\|, \quad \forall \alpha_1, \alpha_2 \in [0, 1] \quad (24)$$

- iii) If Assumption D.2 holds, then $\forall \alpha_1, \alpha_2 \in [0, 1]$, we have

$$\|\nabla L_{\text{Fusion}}(\alpha_1, \Theta^*(\alpha_1)) - \nabla L_{\text{Fusion}}(\alpha_2, \Theta^*(\alpha_2))\| \leq \mathcal{L}_{\text{Fusion}}^* \|\alpha_1 - \alpha_2\| \quad (25)$$

where $\mathcal{L}_{\text{Fusion}}^* \triangleq \frac{D_1 C_{\alpha\Theta}^{\text{Total}}}{\mu_{\text{Total}}} + \frac{C_{\alpha\Theta}^{\text{Total}} \tilde{\mathcal{L}}_{\Theta\Theta}^{\text{Total}}}{\mu_{\text{Total}}^2} C_{\Theta}^{\text{Fusion}}$.

Proof. For part i), we define

$$\begin{aligned} \Gamma_1 &\triangleq \mathcal{M}(\alpha, \Theta) \nabla_{\Theta} L_{\text{Fusion}}(\Theta) - \mathcal{M}(\alpha, \Theta^*(\alpha)) \nabla_{\Theta} L_{\text{Fusion}}(\Theta^*(\alpha)), \\ \Gamma_2 &\triangleq \mathcal{M}(\alpha, \Theta) \{ \nabla_{\Theta} L_{\text{Fusion}}(\Theta) - \nabla_{\Theta} L_{\text{Fusion}}(\Theta^*(\alpha)) \}, \\ \Gamma_3 &\triangleq \nabla_{\Theta} L_{\text{Fusion}}(\Theta^*(\alpha)) \{ \mathcal{M}(\alpha, \Theta) - \mathcal{M}(\alpha, \Theta^*(\alpha)) \}, \\ \Gamma_4 &\triangleq [\nabla_{\Theta}^2 L_{\text{Total}}(\alpha, \Theta)]^{-1} \{ \nabla_{\alpha\Theta}^2 L_{\text{Total}}(\Theta) - \nabla_{\alpha\Theta}^2 L_{\text{Total}}(\Theta^*(\alpha)) \}, \\ \Gamma_5 &\triangleq \nabla_{\alpha\Theta}^2 L_{\text{Total}}(\Theta^*(\alpha)) \left\{ [\nabla_{\Theta}^2 L_{\text{Total}}(\alpha, \Theta)]^{-1} - [\nabla_{\Theta}^2 L_{\text{Total}}(\alpha, \Theta^*(\alpha))]^{-1} \right\}. \end{aligned}$$

Then we have following decomposition by Lemma D.4

$$\begin{aligned}\nabla L_{\text{Fusion}}(\alpha, \Theta^*(\alpha)) - \nabla L_{\text{Fusion}}(\alpha, \Theta) &= \Gamma_1 = \Gamma_2 + \Gamma_3 \\ &= \Gamma_2 + (\Gamma_4 + \Gamma_5) \nabla_{\Theta} L_{\text{Fusion}}(\Theta^*(\alpha))\end{aligned}\quad (26)$$

By the Assumption D.2, we have

$$\begin{aligned}\|\Gamma_2\| &= \|\nabla_{\alpha, \Theta}^2 L_{\text{Total}}(\alpha, \Theta)\| \|\nabla_{\Theta}^2 L_{\text{Total}}(\alpha, \Theta)\|^{-1} \|\nabla_{\Theta} L_{\text{Fusion}}(\Theta) - \nabla L_{\text{Fusion}}(\Theta^*(\alpha))\|, \\ &\leq \frac{C_{\alpha\Theta}^{\text{Total}} \mathcal{L}_{\Theta}^{\text{Fusion}}}{\mu_{\text{Total}}} \|\Theta - \Theta^*(\alpha)\|, \\ \|\Gamma_4\| &\leq \frac{\mathcal{L}_{\alpha\Theta}^{\text{Total}}}{\mu_{\text{Total}}} \|\Theta - \Theta^*(\alpha)\| \\ \|\Gamma_5\| &\leq C_{\alpha\Theta}^{\text{Total}} \frac{\|\nabla_{\Theta}^2 L_{\text{Total}}(\alpha, \Theta(\alpha)) - \nabla_{\Theta}^2 L_{\text{Total}}(\alpha, \Theta^*(\alpha))\|}{\|\nabla_{\Theta}^2 L_{\text{Total}}(\alpha, \Theta(\alpha))\| \|\nabla_{\Theta}^2 L_{\text{Total}}(\alpha, \Theta^*(\alpha))\|}, \\ &\leq \frac{\mathcal{L}_{\Theta\Theta}^{\text{Total}} C_{\alpha\Theta}^{\text{Total}}}{\mu_{\text{Total}}^2} \|\Theta - \Theta^*(\alpha)\|,\end{aligned}$$

where the last inequality holds because given two positive definite symmetric matrices $\mathbf{A}, \mathbf{B} \in \mathcal{R}^{m \times n}$, we have

$$\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\| = \|\mathbf{A}^{-1}(\mathbf{A} - \mathbf{B})\mathbf{B}^{-1}\| \leq \frac{\|\mathbf{A} - \mathbf{B}\|}{\|\mathbf{A}\| \|\mathbf{B}\|} \quad (27)$$

Moreover,

$$\|\nabla_{\Theta} L_{\text{Fusion}}(\Theta^*(\alpha))\| \leq C_{\Theta}^{\text{Fusion}}.$$

Then,

$$\|\nabla L_{\text{Fusion}}(\alpha, \Theta) - \nabla L_{\text{Fusion}}(\alpha, \Theta^*(\alpha))\| \leq \|\Gamma_2\| + (\|\Gamma_4\| + \|\Gamma_5\|) \|\nabla_{\Theta} L_{\text{Fusion}}(\Theta^*(\alpha))\|. \quad (28)$$

We plug the upper bounds of $\|\Gamma_2\|$, $\|\Gamma_4\|$, $\|\Gamma_5\|$ and $\|\nabla_{\Theta} L_{\text{Fusion}}(\Theta^*(\alpha))\|$ into (28), then we finish the proof of part i).

For part ii), by the assumptions, we have:

$$\|\nabla \Theta(\alpha)\| = \|\mathcal{M}(\alpha, \Theta)\| \leq \frac{C_{\alpha\Theta}^{\text{Total}}}{\mu_{\text{Total}}}, \forall \alpha \in [0, 1]. \quad (29)$$

Note that $[0, 1]$ is a closed interval, then part ii) follows immediately.

For part iii), we introduce the similar notations like Γ_1 - Γ_5 to handle $\|\nabla L_{\text{Fusion}}(\alpha_1, \Theta^*(\alpha)) - \nabla L_{\text{Fusion}}(\alpha, \Theta^*(\alpha))\|$, where $\alpha_1, \alpha \in [0, 1]$

$$\begin{aligned}\Omega_1 &= \nabla_{\Theta} L_{\text{Fusion}}(\Theta^*(\alpha_2)) \{\mathcal{M}(\alpha_1, \Theta^*(\alpha_2)) - \mathcal{M}(\alpha_2, \Theta^*(\alpha_2))\}, \\ \Omega_2 &= [\nabla_{\Theta}^2 L_{\text{Total}}(\alpha_1, \Theta^*(\alpha_2))]^{-1} \{\nabla_{\alpha\Theta}^2 L_{\text{Total}}(\Theta^*(\alpha_2)) - \nabla_{\alpha\Theta}^2 L_{\text{Total}}(\Theta^*(\alpha_2))\} = 0, \\ \Omega_3 &= \nabla_{\alpha\Theta}^2 L_{\text{Total}}(\alpha_2, \Theta^*(\alpha_2)) \left\{ [\nabla_{\Theta}^2 L_{\text{Total}}(\alpha_2, \Theta^*(\alpha_2))]^{-1} - [\nabla_{\Theta}^2 L_{\text{Total}}(\alpha_1, \Theta^*(\alpha_2))]^{-1} \right\}.\end{aligned}$$

We have the decomposition:

$$\begin{aligned}\|\nabla L_{\text{Fusion}}(\alpha_2, \Theta^*(\alpha_2)) - \nabla L_{\text{Fusion}}(\alpha_1, \Theta^*(\alpha_2))\| &= \|\Omega_1\| \\ &= \|(\Omega_2 + \Omega_3) \nabla_{\Theta} L_{\text{Fusion}}(\Theta^*(\alpha_2))\| \\ &= \|\Omega_3 \nabla_{\Theta} L_{\text{Fusion}}(\Theta^*(\alpha_2))\| \\ &\leq \|\Omega_3\| \|\nabla_{\Theta} L_{\text{Fusion}}(\Theta^*(\alpha_2))\|,\end{aligned}\quad (30)$$

where the upper bounds for $\|\Omega_3\|$ and $\|\nabla_{\Theta} L_{\text{Fusion}}(\Theta^*(\alpha_2))\|$ can be estimate as following:

$$\begin{aligned}\|\Omega_3\| &\leq \frac{C_{\alpha\Theta}^{\text{Total}} \tilde{\mathcal{L}}_{\Theta\Theta}^{\text{Total}}}{\mu_{\text{Total}}^2} \|\alpha_1 - \alpha_2\| \\ \|\nabla_{\Theta} L_{\text{Fusion}}(\Theta^*(\alpha_2))\| &\leq C_{\Theta}^{\text{Fusion}}\end{aligned}$$

Plug these bounds into (30), we get:

$$\|\nabla L_{\text{Fusion}}(\alpha_2, \Theta^*(\alpha_2)) - \nabla L_{\text{Fusion}}(\alpha_1, \Theta^*(\alpha_2))\| \leq D_2 \|\alpha_1 - \alpha_2\|, \quad (31)$$

where $D_2 \triangleq \frac{C_{\alpha\Theta}^{\text{Total}} \tilde{\mathcal{L}}_{\Theta\Theta}^{\text{Total}}}{\mu_{\text{Total}}^2}$. Now, we can prove the part iii)

$$\begin{aligned}
\|\nabla L_{\text{Fusion}}(\alpha_1, \Theta^*(\alpha_1)) - \nabla L_{\text{Fusion}}(\alpha_2, \Theta^*(\alpha_2))\| &\leq \|\nabla L_{\text{Fusion}}(\alpha_1, \Theta^*(\alpha_1)) - \nabla L_{\text{Fusion}}(\alpha_1, \Theta^*(\alpha_2))\| \\
&\quad + \|\nabla L_{\text{Fusion}}(\alpha_1, \Theta^*(\alpha_2)) - \nabla L_{\text{Fusion}}(\alpha_2, \Theta^*(\alpha_2))\| \\
&\leq D_1 \|\Theta^*(\alpha_1) - \Theta^*(\alpha_2)\| + D_2 \|\alpha_1 - \alpha_2\| \\
&\leq \left(\frac{D_1 C_{\alpha\Theta}^{\text{Total}}}{\mu_{\text{Total}}} + D_2 \right) \|\alpha_1 - \alpha_2\| \\
&= \mathcal{L}_{\text{Fusion}}^* \|\alpha_1 - \alpha_2\|,
\end{aligned} \tag{32}$$

where the third inequality hold because we use the result from the part ii) that $\Theta^*(\alpha)$ is a $\frac{C_{\alpha\Theta}^{\text{Total}}}{\mu_{\text{Total}}}$ -Lipschitz continuous function *w.r.t* α . $\square$

Lemma D.6. (Last-iterate convergence for SGD) Given a μ -strongly convex and L -smooth function $f : \mathcal{R}^d \mapsto \mathcal{R}$ and assume that $\{z_k\}_{k \geq 0}, z_k \in \mathcal{R}^d$ is the sequence generated by the stochastic gradient descent method (SGD) to approximate the minimizer z^* of f under $\mathcal{R}^d$. If stochastic gradient oracle is unbiased and variance bounded by σ . Let we choose the step-size $\eta_k = \min \left\{ \frac{2}{\mu(k+1)}, \frac{1}{L} \right\}$, then we have the following convergence result:

$$\mathbb{E}[\|z_k - z^*\|^2] \leq \frac{4}{k+1} \max \left\{ \frac{\sigma^2}{\mu^2}, \|z_0 - z^*\|^2 \right\}, \quad \forall k \in \mathcal{N}_+ \tag{33}$$

Proof. Recall the SGD update rule and our assumptions:

$$z_{k+1} = z_k - \eta_k g_k, \quad \mathbb{E}[g_k | z_k] = \nabla f(z_k), \quad \mathbb{E}[\|g_k - \nabla f(z_k)\|^2 | z_k] \leq \sigma^2, \tag{34}$$

where g_k is the stochastic gradient at z_k . By the strongly convexity and smoothness of f , we can get:

$$\begin{aligned}
\mathbb{E}[\|z_{k+1} - z^*\|^2 | z_k] &= \mathbb{E}[\|z_k - \eta_k g_k - z^*\|^2 | z_k] \\
&= \|z_k - z^*\|^2 - 2\eta_k \langle \nabla f(z_k), z_k - z^* \rangle + \eta_k^2 \mathbb{E}[\|g_k\|^2 | z_k] \\
&\leq \|z_k - z^*\|^2 - 2\eta_k \left(f(z_k) - f(z^*) + \frac{\mu}{2} \|z_k - z^*\|^2 \right) + \eta_k^2 \mathbb{E}[\|g_k\|^2 | z_k] \\
&\stackrel{(*)}{\leq} \|z_k - z^*\|^2 - 2\eta_k \left(f(z_k) - f(z^*) + \frac{\mu}{2} \|z_k - z^*\|^2 \right) + \eta_k^2 (2L(f(z_k) - f(z^*)) + \sigma^2) \\
&= (1 - \mu\eta_k) \|z_k - z^*\|^2 - 2\eta_k (1 - L\eta_k) (f(z_k) - f(z^*)) + \eta_k^2 \sigma^2 \\
&\leq (1 - \mu\eta_k) \|z_k - z^*\|^2 + \eta_k^2 \sigma^2,
\end{aligned} \tag{35}$$

where the inequality $(*)$ holds because we use $\mathbb{E}[\|g_k\|^2 | z_k] = \mathbb{E}[\|g_k - \nabla f(z_k)\|^2 | z_k] + \mathbb{E}[\|\nabla f(z_k)\|^2] + \mathbb{E}[\langle g_k - \nabla f(z_k), \nabla f(z_k) \rangle | z_k] = \mathbb{E}[\|g_k - \nabla f(z_k)\|^2 | z_k] + \mathbb{E}[\|\nabla f(z_k)\|^2] \leq \sigma^2 + \mathbb{E}[\|\nabla f(z_k)\|^2]$ and a well-known property of the smooth function

$$\|\nabla f(\mathbf{x})\| \leq 2L(f(\mathbf{x}) - f(\mathbf{x}^*)). \tag{36}$$

Then we take the expectation on the both sides, we get

$$\mathbb{E}[\|z_{k+1} - z^*\|^2] \leq (1 - \mu\eta_k) \mathbb{E}[\|z_k - z^*\|^2] + \eta_k^2 \sigma^2 \tag{37}$$

Let $Q = \max \left\{ \frac{4\sigma^2}{\mu^2}, 4\|z_0 - z^*\|^2 \right\}$. We use induction to prove our statement. It is trivial to check that for $k = 0$. Suppose that it is true for some $k > 0$.

i) If $\eta_k = \frac{2}{\mu(k+1)}$

$$\begin{aligned}
\mathbb{E}[\|z_{k+1} - z^*\|^2] &\leq (1 - \mu\eta_k) \mathbb{E}[\|z_k - z^*\|^2] + \eta_k^2 \sigma^2 \\
&\leq \frac{k-1}{k+1} \frac{Q}{k+1} + \frac{4\sigma^2}{\mu^2(k+1)^2} \\
&\leq \frac{k-1}{(k+1)^2} Q + \frac{1}{(k+1)^2} Q \\
&\leq \frac{Q}{k+2} = \frac{4}{k+2} \max \left\{ \frac{\sigma^2}{\mu^2}, \|z_0 - z^*\|^2 \right\}
\end{aligned} \tag{38}$$

ii) If $\eta_k = \frac{1}{L} \Rightarrow \frac{2}{k+1} \geq \frac{\mu}{L}$

$$\begin{aligned}
\mathbb{E}[\|z_{k+1} - z^*\|^2] &\leq (1 - \mu\eta_k)\mathbb{E}[\|z_k - z^*\|^2] + \eta_k^2\sigma^2 \\
&\leq (1 - \frac{\mu}{L})\frac{Q}{k+1} + \frac{\sigma^2}{L^2} \\
&\leq (1 - \frac{2}{k+1})\frac{Q}{k+1} + \frac{4\sigma^2}{\mu^2(k+1)^2} \\
&\leq \frac{Q}{k+2} = \frac{4}{k+2} \max\left\{\frac{\sigma^2}{\mu^2}, \|z_0 - z^*\|^2\right\}
\end{aligned} \tag{39}$$

By the induction, we have finished the proof. $\square$

Lemma D.7. (descent lemma) If Assumption D.2 hold and we set $\eta_k^{\text{out}} \leq \frac{1}{2\mathcal{L}_{\text{Fusion}}^*}$ in Algo. 1, where $\mathcal{L}_{\text{Fusion}}^*$ is defined by the part iii) in Lemma D.5. Then for any $\alpha \in [0, 1]$, we have

$$\begin{aligned}
L_{\text{Fusion}}(\Theta^*(\alpha_{k+1})) - L_{\text{Fusion}}(\Theta^*(\alpha_k)) &\leq \nabla L_{\text{Fusion}}(\alpha_k, \Theta^*(\alpha_k)) \cdot (\alpha - \alpha_k) + \frac{1}{2\eta_k^{\text{out}}}(|\alpha - \alpha_k|^2 - |\alpha - \alpha_{k+1}|^2) \\
&\quad + |\Delta_t| + \delta_k \cdot (\alpha_k - \alpha) + \eta_k^{\text{out}}|\delta_k|^2,
\end{aligned} \tag{40}$$

where $\Delta_k \triangleq \nabla L_{\text{Fusion}}(\alpha_k, \Theta^*(\alpha_k)) - \nabla L_{\text{Fusion}}(\alpha_k, \hat{\Theta}_k)$ and $\delta_k \triangleq \nabla L_{\text{Fusion}}(\alpha_k, \hat{\Theta}_k) - \tilde{\nabla} L_{\text{Fusion}}(\alpha_k, \hat{\Theta}_k, \varrho_k)$.

Proof. Note that $\mathcal{P}_k(x)$ in Algo. 1 is a α_k^{-1} -strongly convex function. By the definition of strongly convexity (see Definition D.2), $\forall \alpha \in [0, 1]$ we have

$$\mathcal{P}_k(\alpha) - \mathcal{P}_k(\alpha_{k+1}) \geq \frac{d\mathcal{P}_k(x)}{dx}(\alpha_{k+1}) \cdot (\alpha - \alpha_{k+1}) + \frac{1}{2\eta_k^{\text{out}}}|\alpha - \alpha_{k+1}|^2 \tag{41}$$

Note that α_{k+1} is a minimizer of $\mathcal{P}_k(x)$ on $[0, 1]$. So by the optimality condition, $\frac{d\mathcal{P}_k(x)}{dx}(\alpha_{k+1}) \cdot (\alpha - \alpha_{k+1}) \geq 0$, which means we can ignore it in (41). Then by the definition of $\mathcal{P}_k(x)$ in Algo. 1, we can get

$$\begin{aligned}
\tilde{\nabla} L_{\text{Fusion}}(\alpha_k, \hat{\Theta}_k, \varrho_k) \cdot (\alpha_{k+1} - \alpha) &\leq \frac{1}{2\eta_k^{\text{out}}}(|\alpha - \alpha_k|^2 - |\alpha - \alpha_{k+1}|^2) \\
&\quad - \frac{1}{2\eta_k^{\text{out}}}|\alpha_{k+1} - \alpha_k|^2
\end{aligned} \tag{42}$$

The part iii) in Lemma D.5 have shown that $L_{\text{Fusion}}(\Theta^*(\alpha))$ is a $\mathcal{L}_{\text{Fusion}}^*$ -smooth function w.r.t to α , so we can use the Proposition D.1

$$\begin{aligned}
L_{\text{Fusion}}(\Theta^*(\alpha_{k+1})) - L_{\text{Fusion}}(\Theta^*(\alpha_k)) &\leq \nabla L_{\text{Fusion}}(\alpha_k, \Theta^*(\alpha_k)) \cdot (\alpha_{k+1} - \alpha_k) + \frac{\mathcal{L}_{\text{Fusion}}^*}{2}|\alpha_{k+1} - \alpha_k|^2 \\
&= \nabla L_{\text{Fusion}}(\alpha_k, \Theta^*(\alpha_k)) \cdot (\alpha_{k+1} - \alpha) + \nabla L_{\text{Fusion}}(\alpha_k, \Theta^*(\alpha_k)) \cdot (\alpha - \alpha_k) \\
&\quad + \frac{\mathcal{L}_{\text{Fusion}}^*}{2}|\alpha_{k+1} - \alpha_k|^2 \\
&= \nabla L_{\text{Fusion}}(\alpha_k, \Theta^*(\alpha_k)) \cdot \alpha - \alpha_k + \tilde{\nabla} L_{\text{Fusion}}(\alpha_k, \hat{\Theta}_k, \varrho_k) \cdot (\alpha_{k+1} - \alpha) \\
&\quad + \tilde{\Delta}_k \cdot (\alpha_{k+1} - \alpha) + \frac{\mathcal{L}_{\text{Fusion}}^*}{2}|\alpha_{k+1} - \alpha_k|^2 \\
&\stackrel{(42)}{\leq} \nabla L_{\text{Fusion}}(\alpha_k, \Theta^*(\alpha_k)) \cdot (\alpha - \alpha_k) + \frac{1}{2\eta_k^{\text{out}}}(|\alpha - \alpha_k|^2 - |\alpha - \alpha_{k+1}|^2) \\
&\quad + \left(\frac{\mathcal{L}_{\text{Fusion}}^*}{2} - \frac{1}{2\eta_k^{\text{out}}}\right)|\alpha_{k+1} - \alpha_k|^2 + \tilde{\Delta}_k \cdot (\alpha_{k+1} - \alpha),
\end{aligned} \tag{43}$$

where $\tilde{\Delta}_k \triangleq \nabla L_{\text{Fusion}}(\alpha_k, \Theta^*(\alpha_k)) - \tilde{\nabla} L_{\text{Fusion}}(\alpha_k, \hat{\Theta}_k, \varrho_k)$ and we notice that

$$\begin{aligned}
\tilde{\Delta}_k \cdot (\alpha_{k+1} - \alpha) &= \Delta_k \cdot (\alpha_{k+1} - \alpha) + \delta_k \cdot (\alpha_{k+1} - \alpha_k) + \delta_k \cdot (\alpha_k - \alpha) \\
&\leq |\Delta_t| + \delta_k \cdot (\alpha_k - \alpha) + \eta_k^{\text{out}}|\delta_k|^2 + \frac{1}{4\eta_k^{\text{out}}}|\alpha_{k+1} - \alpha_k|^2,
\end{aligned} \tag{44}$$

where the inequality hold because we use $\alpha_k, \alpha \in [0, 1]$, Cauchy-Schwarz inequality and Young's inequality. Plug this upper bound into (43) and note that we assume that $\eta_k^{\text{out}} \leq \frac{1}{2\mathcal{L}_{\text{Fusion}}^*}$, then we finish the proof. $\square$

Theorem D.8 (Formal). Let Assumption D.2 and Assumption D.3 hold. We first fix the number of outer iterations in Algo. 1 as T and set $\eta_t^{\text{in}} = \min\left\{\frac{2}{\mu_{\text{Total}}(t+1)}, \frac{1}{\mathcal{L}_{\Theta}^{\text{Total}}}\right\}$, $t_k = T - 1$ and $\eta_k^{\text{out}} = \frac{1}{2\mathcal{L}_{\text{Fusion}}^*\sqrt{T}}$, where $\mathcal{L}_{\text{Fusion}}^*\sqrt{T}$ is defined by the part iii) in Lemma

D.5. Let L_{Fusion}^* is the optimal solution corresponding with the $\alpha^* \in [0, 1]$ and $\bar{\alpha}_T \triangleq \frac{\sum_{t=1}^T \alpha_t}{T}$. If L_{Fusion} is convex w.r.t $\alpha \in [0, 1]$, we have following convergence result for Algo. 1.

$$\mathbb{E}[L_{\text{Fusion}}(\Theta^*(\bar{\alpha}_T))] - L_{\text{Fusion}}^* \leq \frac{1}{\sqrt{T}} \left[\mathcal{L}_{\text{Fusion}}^* |\alpha_0 - \alpha^*|^2 + 2C_1 D_1 + \frac{\lambda^2}{2\mathcal{L}_{\text{Fusion}}^*} \right], \quad (45)$$

where C_1 , D_1 and λ_1 are all constants that only dependent on the problem parameters

$$\begin{aligned} C_1 &= \max \left\{ \frac{\sigma_{\Theta}^{\text{Total}}}{\mu_{\text{Total}}}, \arg \max_{\alpha \in [0,1]} \|\hat{\Theta}_0 - \Theta^*(\alpha)\| \right\}, \\ D_1 &= \frac{C_{\alpha\Theta}^{\text{Total}} \mathcal{L}_{\Theta}^{\text{Fusion}}}{\mu_{\text{Total}}} + \left(\frac{\mathcal{L}_{\alpha\Theta}^{\text{Total}}}{\mu_{\text{Total}}} + \frac{\mathcal{L}_{\Theta\Theta}^{\text{Total}} C_{\alpha\Theta}^{\text{Total}}}{\mu_{\text{Total}}^2} \right) C_{\Theta}^{\text{Fusion}}, \\ \lambda^2 &= 16((C_{\Theta}^{\text{Fusion}})^2 + (\sigma_{\Theta}^{\text{Fusion}}))((\sigma_{\alpha\Theta}^{\text{Total}})^2 + (C_{\alpha\Theta}^{\text{Total}})^2) \left(\frac{1}{\mu_{\text{Total}}^2} + \sigma_{\mathbf{H}}^2 \right) + \frac{2(C_{\alpha\Theta}^{\text{Total}})^2 (C_{\Theta}^{\text{Fusion}})^2}{\mu_{\text{Total}}^2}. \end{aligned} \quad (46)$$

Proof. If we let $\mathcal{F}_k = \sigma(\alpha_0, \dots, \alpha_k)$, then we have following observations:

$$\begin{cases} \sigma(\Delta_k), \sigma(\hat{\Theta}_k) \subseteq \mathcal{F}_k \cup \sigma(\zeta_{[t_k]}^{(1)}) \\ \sigma(\alpha_k) \subseteq \mathcal{F}_k \\ \sigma(\delta_k) \subseteq \mathcal{F}_k \cup \sigma(\varrho_k) = \mathcal{F}_{k+1} \end{cases} \quad (47)$$

Using these σ -algebras we define, we can analysis the following conditional expectations:

$$\begin{aligned} \mathbb{E}[\mathcal{H}_{\Theta\Theta}(\alpha_k, \hat{\Theta}_k, \zeta_k^{(3)}) | \mathcal{F}_k \cup \sigma(\zeta_{[t_k]}^{(1)})] &= \mathbb{E}[\mathcal{H}_{\Theta\Theta}(\alpha_k, \hat{\Theta}_k, \zeta_k^{(3)}) | \alpha_k, \hat{\Theta}_k, \mathcal{F}_k \cup \sigma(\zeta_{[t_k]}^{(1)})] \\ &= \mathbb{E}[\mathcal{H}_{\Theta\Theta}(\alpha_k, \hat{\Theta}_k, \zeta_k^{(3)}) | \alpha_k, \hat{\Theta}_k] \\ &\stackrel{a.m.s}{=} [\nabla_{\Theta}^2 L_{\text{Total}}(\alpha_k, \hat{\Theta})]^{-1} \end{aligned} \quad (48)$$

$$\begin{aligned} \mathbb{E}[\delta_k | \mathcal{F}_k \cup \sigma(\zeta_{[t_k]}^{(1)})] &= \nabla_{\alpha\Theta}^2 L_{\text{Total}}(\alpha_k, \hat{\Theta}_k) \left\{ \mathbb{E}[\mathcal{H}_{\Theta\Theta}(\alpha_k, \hat{\Theta}_k, \zeta_k^{(3)}) | \mathcal{F}_k \cup \sigma(\zeta_{[t_k]}^{(1)})] - [\nabla_{\Theta}^2 L_{\text{Total}}(\alpha_k, \hat{\Theta})]^{-1} \right\} \\ &\quad \cdot \nabla_{\Theta} L_{\text{Fusion}}(\hat{\Theta}_k) \stackrel{a.m.s}{=} 0 \end{aligned} \quad (49)$$

$$\begin{aligned} \mathbb{E}[\|\tilde{\nabla}_{\alpha\Theta}^2 G_{\text{Total}}(\alpha_k, \hat{\Theta}_k, \zeta_k^{(2)})\|^2 | \mathcal{F}_k \cup \sigma(\zeta_{[t_k]}^{(1)})] &= \mathbb{E}[\|\tilde{\nabla}_{\alpha\Theta}^2 G_{\text{Total}}(\alpha_k, \hat{\Theta}_k, \zeta_k^{(2)})\|^2 | \alpha_k, \hat{\Theta}_k] \\ &\leq 2\mathbb{E}[\|\tilde{\nabla}_{\alpha\Theta}^2 G_{\text{Total}}(\alpha_k, \hat{\Theta}_k, \zeta_k^{(2)}) - \nabla_{\alpha\Theta}^2 L_{\text{Total}}(\alpha_k, \hat{\Theta}_k)\|^2 | \alpha_k, \hat{\Theta}_k] \\ &\quad + 2\|\nabla_{\alpha\Theta}^2 L_{\text{Total}}(\alpha_k, \hat{\Theta}_k)\|^2 \\ &\leq 2(\sigma_{\alpha\Theta}^{\text{Total}})^2 + 2(C_{\alpha\Theta}^{\text{Total}})^2 \end{aligned} \quad (50)$$

$$\mathbb{E}[\|\mathcal{H}_{\Theta\Theta}(\alpha_k, \hat{\Theta}, \zeta_k^{(3)})\|^2 | \mathcal{F}_k \cup \sigma(\zeta_{[t_k]}^{(1)})] \leq \frac{2}{\mu_{\text{Total}}^2} + 2\sigma_{\mathbf{H}}^2 \quad (51)$$

Therefore, we can get:

$$\begin{aligned} \mathbb{E}[|\delta_k|^2] &= \mathbb{E}[\mathbb{E}[|\delta_k|^2 | \mathcal{F}_k \cup \sigma(\zeta_{[t_k]}^{(1)})]] \leq 16((C_{\Theta}^{\text{Fusion}})^2 + \sigma_{\Theta}^{\text{Fusion}})((\sigma_{\alpha\Theta}^{\text{Total}})^2 + (C_{\alpha\Theta}^{\text{Total}})^2) \left(\frac{1}{\mu_{\text{Total}}^2} + \sigma_{\mathcal{H}}^2 \right) \\ &\quad + \frac{2(C_{\alpha\Theta}^{\text{Total}})^2 (C_{\Theta}^{\text{Fusion}})^2}{\mu_{\text{Total}}^2} \\ &:= \lambda^2 \end{aligned} \quad (52)$$

Note that $L_{\text{Total}}(\alpha, \Theta)$ is a μ_{Total} -strongly convex and $\mathcal{L}_{\Theta}^{\text{Total}}$ -smooth function with respect to Θ . If we let $\eta_t^{\text{in}} = \min \left\{ \frac{2}{\mu_{\text{Total}}(t+1)}, \frac{1}{\mathcal{L}_{\Theta}^{\text{Total}}} \right\}$. By the Lemma D.6, we can get:

$$\begin{aligned} \mathbb{E}[\|\hat{\Theta}_k(\alpha_k) - \Theta^*(\alpha_k)\|] &= \mathbb{E}[\mathbb{E}_{\zeta_{[t_k]}^{(1)}} [\|\hat{\Theta}_k(\alpha_k, \zeta_{[t_k]}^{(1)}) - \Theta^*(\alpha_k)\| | \alpha_k, \mathcal{F}_k]] \\ &= \mathbb{E}[\mathbb{E}_{\zeta_{[t_k]}^{(1)}} [\|\hat{\Theta}_k(\alpha_k, \zeta_{[t_k]}^{(1)}) - \Theta^*(\alpha_k)\| | \alpha_k]] \\ &\leq \frac{2}{\sqrt{t_k+1}} \mathbb{E} \left[\max \left\{ \frac{\sigma_{\Theta}^{\text{Total}}}{\mu_{\text{Total}}}, \|\hat{\Theta}_0 - \Theta^*(\alpha_k)\| \right\} \right] \\ &\leq \frac{2}{\sqrt{t_k+1}} \max \left\{ \frac{\sigma_{\Theta}^{\text{Total}}}{\mu_{\text{Total}}}, \arg \max_{\alpha \in [0,1]} \|\hat{\Theta}_0 - \Theta^*(\alpha)\| \right\} \\ &:= A_k \end{aligned} \quad (53)$$

Note that we have shown that $\Theta^*(\alpha)$ is Lipschitz continuous function with respect to α so $\arg \max_{\alpha \in [0,1]} \|\hat{\Theta}_0 - \Theta^*(\alpha)\| \leq +\infty$. Then, we can estimate the upper bound of $\mathbb{E}[\|\Delta_k\|]$:

$$\begin{aligned} |\Delta_k| &= \mathbb{E}[\|\nabla L_{\text{Fusion}}(\alpha_k, \Theta^*(\alpha_k)) - \nabla L_{\text{Fusion}}(\alpha_k, \hat{\Theta}_k)\|] \leq D_1 \mathbb{E}[\|\Theta^*(\alpha_k) - \hat{\Theta}_k\|] \\ &\leq D_1 A_k, \end{aligned} \quad (54)$$

where the first hold because the part i) of Lemma D.5. Now, we substitute α^* into (40) and then take expectation on the both sides of (40) and plug (52) and (54) into it.

$$\begin{aligned} \mathbb{E}[L_{\text{Fusion}}(\Theta^*(\alpha_{k+1}))] - \mathbb{E}[L_{\text{Fusion}}(\Theta^*(\alpha_k))] &\leq \mathbb{E}[\nabla L_{\text{Fusion}}(\alpha_k, \Theta^*(\alpha_k)) \cdot (\alpha - \alpha_k)] \\ &\quad + \frac{1}{2\eta_k^{\text{out}}} \mathbb{E}[|\alpha^* - \alpha_k|^2 - |\alpha^* - \alpha_{k+1}|^2] \\ &\quad + D_1 A_k + \lambda^2 \eta_k^{\text{out}} \end{aligned} \quad (55)$$

If $L_{\text{Fusion}}(\Theta^*(\alpha))$ is convex w.r.t α , then $\mathbb{E}[\nabla L_{\text{Fusion}}(\alpha_k, \Theta^*(\alpha_k)) \cdot (\alpha^* - \alpha_k)] \leq L_{\text{Fusion}}^* - \mathbb{E}[L_{\text{Fusion}}(\Theta^*(\alpha_k))]$. Therefore, we have:

$$\begin{aligned} \mathbb{E}[L_{\text{Fusion}}(\Theta^*(\alpha_k))] - L_{\text{Fusion}}^* &\leq \frac{1}{2\eta_{k-1}^{\text{out}}} \mathbb{E}[|\alpha^* - \alpha_{k-1}|^2 - |\alpha^* - \alpha_k|^2] \\ &\quad + D_1 A_{k-1} + \lambda^2 \eta_{k-1}^{\text{out}} \end{aligned} \quad (56)$$

Then take summation from $t = 1, \dots, T$ on the both sides, we can get:

$$\sum_{t=1}^T \mathbb{E}[L_{\text{Fusion}}(\Theta^*(\alpha_k))] - T L_{\text{Fusion}}^* \leq 2\mathcal{L}_{\text{Fusion}}^* \sqrt{T} |\alpha_0 - \alpha^*|^2 + D_1 \sum_{t=1}^T A_{k-1} + \lambda^2 \frac{\sqrt{T}}{2\mathcal{L}_{\text{Fusion}}^*} \quad (57)$$

Note that $\sum_{t=1}^T A_{k-1} = 2C_1 \sqrt{T}$ and $\sum_{t=1}^T \mathbb{E}[L_{\text{Fusion}}(\Theta^*(\alpha_T))] = \sum_{t=1}^T \mathbb{E}[L_{\text{Fusion}}(\Theta^*(\frac{1}{T} \sum_{k=1}^T \alpha_k))]$ $\leq \frac{1}{T} \sum_{t=1}^T \mathbb{E}[L_{\text{Fusion}}(\Theta^*(\alpha_k))]$. Then we multiply T on the both sides, we finish the proof. $\square$

REFERENCES

- [1] Y. Fan, W. Xu, H. Wang, J. Wang, and S. Guo, "PMR: prototypical modal rebalance for multimodal learning," in *CVPR*, 2023, pp. 20 029–20 038.
- [2] Y. Wei and D. Hu, "Mmpareto: Boosting multimodal learning with innocent unimodal assistance," in *ICML*, 2024, pp. 52 559–52 572.
- [3] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *CVPR*, 2022, pp. 8228–8237.
- [4] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP*, 2020, pp. 721–725.
- [5] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Python in Science Conference*, 2015, pp. 18–24.
- [6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019, pp. 4171–4186.
- [7] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in twitter with hierarchical fusion model," in *ACL*, 2019, pp. 2506–2515.
- [8] J. Yu and J. Jiang, "Adapting BERT for target-oriented multimodal sentiment classification," in *IJCAI*, 2019, pp. 5408–5414.
- [9] L. H. Loomis and S. Shlomo, *Advanced Calculus*, 1990, pp. 164–169.