Appendix

7 Formalizing Cross-modal Generalization

In the main text, we have established, with some degree of formality, a framework for studying cross-modal generalization. In this section, we extend and add in details surrounding the discussion in the main text. This will allow us to understanding our method in a more precise manner.

Cross-modal generalization is a learning paradigm to quickly perform new tasks in a target modality despite being trained on a different source modality. To formalize this paradigm, we build on the definition of meta-learning [25] and generalize it to study multiple input modalities. The goal of meta-learning can be broadly defined as using labeled data for existing source tasks to learn representations that enable fast learning on unseen target tasks [33]. To reason over multiple modalities and tasks, we start by defining M different heterogeneous input spaces (modalities) and N different label spaces (tasks). We denote a modality by an index $m \in \{1, ..., M\}$ and a task by $n \in \{1, ..., N\}$.

Each classification problem $\mathcal{T}(m, n)$ is defined as a triplet with a modality, task, plus a joint distribution: $\mathcal{T}(m, n) = (\mathcal{X}_m, \mathcal{Y}_n, p_{m,n}(x, y))$. $\mathcal{X}_m$ denotes the input space and $\mathcal{Y}_n$ the label space sampled from a distribution $p(m, n) := p(\mathcal{X}_m, \mathcal{Y}_n)$ given by a marginal over the entire *meta-distribution*, $p(x_1, ..., x_M, y_1, ...y_N, \mathcal{X}_{m_1}, ...\mathcal{X}_{m_M}, \mathcal{Y}_{n_1}, ...\mathcal{Y}_{n_N})$. The meta-distribution gives the underlying relationships between all modalities and tasks through a hierarchical generative process $m_i \sim p(m), n_j \sim p(n)$: first picking a modality and task (m_i, n_j) from priors $p(m)$ and $p(n)$ over input and output spaces, before drawing data x_i from $\mathcal{X}_{m_i}$ and labels y_j from $\mathcal{Y}_{n_j}$. Within each classification problem is also an underlying pairing function mapping inputs to labels through $p_{m,n}(x, y) := p(x, y|m, n)$ for all $x \in \mathcal{X}_m,\ y \in \mathcal{Y}_n$ representing the true data labeling process. Note that in practice $p_{m,n}(x, y)$ is never known but instead represented as (modality, label) pairs as collected and annotated as real-world datasets.

To account for generalization over modalities and tasks, cross-modal generalization involves learning a single function f_w with parameters w over the meta-distribution with the following objective:

Definition 3. *Cross-modal generalization is a maximization problem given by*

$$\arg\max_w \mathcal{L}[f_w] := \arg\max_w \mathop{\mathbb{E}}_{\substack{m,n\sim p(m,n) \\ x,y\sim p_{m,n}(x,y)}} \log\left[\frac{f_w(x, y, m, n)}{p(x, y|m, n)}\right]. \tag{3}$$

When $p(n)$ is a delta distribution, we say that the problem is single task; otherwise, it is multi-task. $p(m)$ is any arbitrary distribution over the source domains.

We call eq (3) the generalization loss and the goal of any model we consider is to minimize this loss. Notice that this loss is lower bounded by 0, and is achievable when $f_w(x, y, m, n) = p(x, y|m, n)$. A model f_w that achieves 0 loss in eq (3) is said to achieve perfect generalization.

7.1 Cross-modal Few-shot learning

In practice, the space between modalities and tasks is only *partially observed*: $p(x, y|m, n)$ is only observed for certain modalities and tasks (e.g. labeled classification tasks for images [11], or paired data across image, text, and audio in online videos [1]). For other modality-task pairs, we can only obtain inaccurate estimates $q(x, y|m, n)$, often due to having only *limited labeled data*. We are now ready to define a *few-shot learning* problem.

Definition 4. *Let $\mathcal{M}$ be a subset of all the possible pairings of modality and task spaces. A meta-learning problem is said to be (partially) low resource if for all $m,\ n \in \mathcal{M}$, $p(x, y|m, n)$ is not known exactly, and has to be estimated using $q(x, y|m, n) \neq p(x, y|m, n)$.*

Therefore, the subset $\mathcal{M}$ can be called the low-resource subset, and any task associated with $\mathcal{M}$ is a low-resource task. Note that this definition is equivalent to a situation where we have infinitely many data points for the high resource tasks, and a finite number of data points for the low-resource tasks. In practice, $q(x, y|m, n)$ is an (imperfect) estimation of $p(x, y|m, n)$ due to limited labeled

data. Mathematically, for a few-shot meta-learning problem, the optimization objective becomes

$$\arg\max_w \mathcal{L}_q[f_w] := \arg\max_w \mathop{\mathbb{E}}_{\substack{m,n \\ \sim p(m,n)}} \Bigg\{ \underbrace{\mathbf{1}_{(m,n)\notin\mathcal{M}} \mathop{\mathbb{E}}_{\substack{x,y \\ \sim p_{m,n}(x,y)}} \log\left[\frac{f_w(x,y,m,n)}{p(x,y|m,n)}\right]}_{\text{high resource subset}} + \underbrace{\mathbf{1}_{(m,n)\in\mathcal{M}} \mathop{\mathbb{E}}_{\substack{x,y \\ \sim q_{m,n}(x,y)}} \log\left[\frac{f_w(x,y,m,n)}{q(x,y|m,n)}\right]}_{\text{low resource subset}} \Bigg\}, \tag{4}$$

where $\mathbf{1}$ is the indicator function. This new optimization objective no longer matches the generalization objective $\mathcal{L}$ in eq (3). The minimizer of this equation is $f_w(x,y,m,n) = q(x,y|m,n)$, which has an generalization error $\Pr\{(m,n) \in \mathcal{M}\}\mathrm{KL}(p;q)$, where $\mathrm{KL}(\cdot;\cdot)$ is the KL-divergence measuring how inaccurate the real-life estimates q are due to limited labeled data.

What is the minimal extra supervision required to perform cross-modal generalization under only partial observability? To answer this question, we first define the minimum requirements on observed data, which we call the *minimum visibility assumption*:

Assumption 1. *(Minimum visibility) For every task n, there is at least one domain m such that $p(x,y|m,n)$ is known. Likewise, for every domain m, there is at least one task n such $p(x,y|m,n)$ is known. All the single variable marginal distributions $p(x)$, $p(y)$ are also known.*

In practice, we say that a distribution is known if it can be accurately estimated. This is the minimum assumption required to ensure that all modalities and tasks are accessible. It is helpful to think about this *partial observability* as a bipartite graph $G = (V_x, V_y, E)$ between a modality set V_x and task set V_y (see Figure 8). A solid directed edge from $u \in V_x$ to $v \in V_y$ represents learning a classifier from modality u for task v given an abundance of observed labeled data, which incurs negligible generalization error. Since it is unlikely for all edges between V_x and V_y to exist, define the *low-resource subset* $\mathcal{M}$ as the complement of E in $V_x \times V_y$. $\mathcal{M}$ represents the set of low-resource modalities and tasks where it is difficult to obtain labeled data. The focus of cross-modal generalization is to learn a classifier in $\mathcal{M}$ as denoted by a dashed edge. In contrast to solid edges, the lack of data in $\mathcal{M}$ incurs large error along dashed edges. It is helpful to differentiate solid vs dashed edges by writing them as weighted edges (u, v, ϵ), where ϵ denotes error incurred. The visibility assumption says that there is at least a solid in/out edge for every vertex in V_m and V_n.

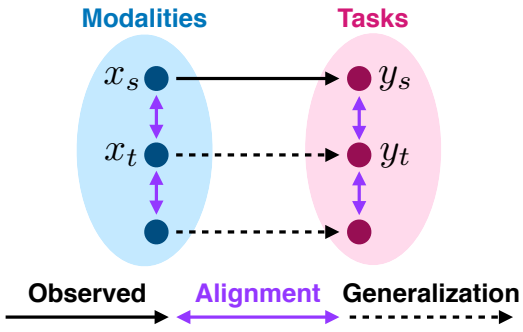

Figure 8: A modality-task graph contains a subset of observed edges through labeled datasets for specific modalities and tasks. Generalizing to the remaining modalities and tasks (dotted edge) requires bridging modalities through alignment.

7.2 Cross-modal Alignment

Therefore, the challenge in cross-modal generalization amounts to finding the path of lowest cumulative error between an input target modality $x_t \in V_x$ and output task $y_t \in V_y$ in $\mathcal{M}$. The key insight is to leverage *cross-modal information* to "bridge" modalities that are each labeled for only a subset of tasks (see purple edges in Figure 8). We model cross-modal information as $p(x_s, x_t)$, i.e. *alignment* between modalities x_s and x_t, where x_s is a source modality with high-resource data and labels (x_s, y_s). When there is an abundance of paired data (x_s, x_t) (solid purple edge), we say that *strong* alignment exists; otherwise, only *weak* alignment exits. Since strong alignment incurs negligible error in estimating $p(x_s, x_t)$, the alternative *cross-modal path* $P = \{(x_t, x_s), (x_s, y_s), (y_s, y_t)\}$ might link x_t and y_t with *lower* weighted error and is preferable to direct low-resource training for the dashed edge (x_t, y_t). When only weak alignment is available, a trade-off emerges and one has to choose between the error induced by direct low-resource training and the error induced by weak alignment. (y_s, y_t) models relationships across source and target tasks using approaches such as multi-task [7] or meta-learning [17]. More formally,

Definition 5. *Let $p(x_i, x_j)$ be known for $x_i \in \mathcal{D}^x_{m_i}$, $x_j \in \mathcal{D}^x_{m_j}$ and $i \neq j$. If both $p(x_i|x_j)$ and $p(x_j|x_i)$ are delta distributions, i.e., if there is a one-to-one mapping between x_i and x_j, we say that there is a strong alignment between modality m_i and m_j. Otherwise, there is only weak alignment.*

We now show that strong alignment across modalities can achieve optimal generalization error for tasks in the low-resource subset $\mathcal{M}$.

Proposition 2. *(Benefit of strong alignment). Let all the modalities be pairwise strongly-aligned, then we can define a surrogate loss function $\tilde{\mathcal{L}}[f_w]$ such that $\mathcal{L}[\arg\min_{f_w} \tilde{\mathcal{L}}[f_w]] = 0$.*

Proof. Let $\mathcal{T}_{s,t}$ be in the low-resource set, where we only know $q(x_t, y)$. We want to show that we can recover $p(x_t, y)$ from alignment information. By the assumption of visibility, for task t, there is a strongly aligned modality $s \neq t$ for which we know $p(x_s, x_t)$. By Bayes' rule $p(x_t, x_s, y) = p(x_t|x_s, y)p(x_s, y)$, but x_s is conditionally independent of y if x_t is known due to the existence of one-to-one mapping between them. Therefore, we can calculate $p(x_t, x_s, y) = p(x_t|x_s)p(x_s, y)$ recover the desired label $p(x_t, y) = \int p(x_s, x_t, y)\, dx_s$. Now we can replace $q(x_t, y)$ by the recovered $p(x_t, y)$ in the loss function, thus achieving perfect generalization on this task. □

This implies that if strong alignment is achievable, then one can achieve perfect generalization in the low-resource subset $\mathcal{M}$. We also note that a key property we used in the proof is that $p(x_t|x_s) = p(x_t|x_s, y)$. For weak alignment, this property does not hold and perfect generalization is no longer achievable, and one needs to tradeoff the error induced by weak alignment with the error from minimizing q directly (i.e. few-shot supervised learning). We further explain and qualitatively analyze this trade-off in Appendix 11.

Therefore, unlabeled cross-modal information $p(x_s, x_t)$ allows us to bridge modalities that are each labeled for only a subset of tasks and achieve cross-modal generalization to new modalities and tasks in $\mathcal{M}$. In practice, however, $p(x_s, x_t)$ is unknown and needs to be estimated from data, and is the basis for our proposed CROMA approach to estimate $p(x_s, x_t)$ from data and meta-learning to model (y_s, y_t).

7.3 Concerning Weak Alignment

For weak alignment, this property may not hold and perfect generalization may not be achievable. Therefore, one needs to tradeoff the error induced by weak alignment with the error from minimizing q directly (i.e. few-shot supervised learning). This does not necessarily mean that weak alignment will hurt generalization: if $p(x_t|x_s, y) = p(x_t|x_s)$ holds, then perfect generalization can still be achieved. Of course, one might differentiate between the *perfect weak alignment* problem, where the statement $p(x_t|x_s, y) = p(x_t|x_s)$ holds (or, requiring one additional assumption) and *proper weak alignment*, where it does not. One can therefore prove the following corollary.

Corollary 1. *Assuming perfect weak alignment, one can achieve perfect generalization error.*

The proof follows directly from Proposition 1.

8 Experimental Details

The code for running our experiments can be found in the supplementary material. We also provide some experimental details below. Since there are no established benchmarks in cross-modal generalization, we create our own by merging and preprocessing several multimodal datasets. We believe that these two benchmarks for assessing cross-modal generalization (image to audio and text to speech) will also be useful to the broader research community and hence we also open-source all data and data processing code.

8.1 Text to Image

Data: We use the Yummly-28K dataset [43] which contains parallel text descriptions and images of recipes. We create classification labels from the metadata by concatenating the meal type and cuisine, yielding 44 distinct classes. The large number of recipes and shared concepts between text and image makes it an ideal testbed for cross-modal generalization. We used a ResNet pretrained on ImageNet [11] to encode the images, pretrained BERT encoder [12] for text, and a shared network for prediction.

Hyperparameters: We show the hyperparameters used in Table 3.

Table 3: Table of hyperparameters for generalization experiments on text to image task. Batchsize $4/8/16$ indicates the batchsize used for $1/5/10$-shot experiments respectively.

Model	Parameter	Value
Text Encoder	Shared layer size	256
	Batchsize	$4/8/16$
	Activation	ReLU
	Meta Optimizer	SGD
	Optimizer	Adam
	Meta Learning rate	$1e-1$
	Align Learning rate	$1e-3$
	Classifier Learning rate	$1e-3$
	Iterations	800
	Number of evaluation tasks	16
	Loss Margin	0.1
	Width Factor	1.3
	Number of Layers	4
	Blocks Per Layer	$4, 5, 24, 3$

Model	Parameter	Value
Image Encoder	Shared layer size	256
	Batchsize	$4/8/16$
	Activation	ReLU
	Meta Optimizer	SGD
	Optimizer	Adam
	Meta Learning rate	$1e-1$
	Align Learning rate	$1e-3$
	Classifier Learning rate	$1e-3$
	Iterations	800
	Number of evaluation tasks	16
	Loss Margin	0.1
	Intermediate Pooling Function	Max
	Final Pooling Function	Average
	Stride	1

Model	Parameter	Value
Image Classifier	Num hidden layers	1
	Hidden layer size	256
	Batchsize	$4/8/16$
	Activation	ReLU
	Meta Optimizer	SGD
	Optimizer	Adam
	Meta Learning rate	$1e-1$
	Align Learning rate	$1e-3$
	Classifier Learning rate	$1e-3$
	Iterations	800
	Number of evaluation tasks	16
	Loss	MSE
	Teacher forcing rate	0.5

8.2 Image to Audio

Data: To construct our generalization dataset, we combine 100 classes from CIFAR-100 and 10 classes from CIFAR-10 [35] to form 110 image classes, as well as 50 audio classes from ESC-50 [49]. The tasks across these modalities are different (i.e. different classification problems) which requires cross-modal generalization. To bridge these two modalities with partially related label spaces, we define 17 shared classes across the 2 datasets for weak concept alignment. We show the 17 clustered concepts we used for weak alignment in Figure 9. These clusters are obtained by mapping similar classes between the datasets using similarities from WordNet [42] and text cooccurrence. The number

Table 4: Table of hyperparameters for generalization experiments on image to audio task. Batchsize $4/8/16$ indicates the batchsize used for $1/5/10$-shot experiments respectively.

Model	Parameter	Value
Image Encoder	Shared layer size	256
	Batchsize	$4/8/16$
	Activation	ReLU
	Meta Optimizer	SGD
	Optimizer	Adam
	Meta Learning rate	$1e-1$
	Align Learning rate	$1e-3$
	Classifier Learning rate	$1e-3$
	Iterations	800
	Number of evaluation tasks	16
	Loss Margin	0.1
	Width Factor	1.3
	Number of Layers	4
	Blocks Per Layer	$4, 5, 24, 3$

Model	Parameter	Value
Audio Encoder	Shared layer size	256
	Batchsize	$4/8/16$
	Activation	ReLU
	Meta Optimizer	SGD
	Optimizer	Adam
	Meta Learning rate	$1e-1$
	Align Learning rate	$1e-3$
	Classifier Learning rate	$1e-3$
	Iterations	800
	Number of evaluation tasks	16
	Loss Margin	0.1
	Intermediate Pooling Function	Max
	Final Pooling Function	Average
	Stride	1

Model	Parameter	Value
Audio Classifier	Num hidden layers	1
	Hidden layer size	256
	Batchsize	$4/8/16$
	Activation	ReLU
	Meta Optimizer	SGD
	Optimizer	Adam
	Meta Learning rate	$1e-1$
	Align Learning rate	$1e-3$
	Classifier Learning rate	$1e-3$
	Iterations	800
	Number of evaluation tasks	16
	Loss	MSE
	Teacher forcing rate	0.5

of shared classes in train, val, and test, respectively is 12, 8, and 9, and the number of samples is 920, 580, 580, respectively.

Hyperparameters: We show the hyperparameters used in Table 4.

8.3 Text to Speech

Data: The dataset is composed of paired text-speech data from a 99-language subset of the Wilderness dataset [5]. The dataset was collected using text and speech from the Bible. We preprocessed the data so that every language corresponded to a different set of chapters, maximizing the independence

Table 5: Table of hyperparameters for generalization experiments on text to speech task. Batchsize $4/8/16$ indicates the batchsize used for $1/5/10$-shot experiments respectively.

Model	Parameter	Value
Text Encoder	Bidirectional	True
	Embedding dim	256
	Num hidden layers	1
	Hidden layer size	256
	Batchsize	$4/8/16$
	Activation	ReLU
	Meta Optimizer	SGD
	Optimizer	Adam
	Meta Learning rate	$1e-1$
	Align Learning rate	$1e-3$
	Classifier Learning rate	$1e-4$
	Iterations	800
	Loss Margin	0.1
	Number of evaluation tasks	16

Model	Parameter	Value
Speech Encoder	Embedding dim	40
	Num hidden layers	2
	Hidden layer size	256
	Batchsize	$4/8/16$
	Activation	ReLU
	Meta Optimizer	SGD
	Optimizer	Adam
	Meta Learning rate	$1e-1$
	Align Learning rate	$1e-3$
	Classifier Learning rate	$1e-4$
	Iterations	800
	Loss Margin	0.1
	Number of evaluation tasks	16

Model	Parameter	Value
Speech Classifier	Num hidden layers	1
	Hidden layer size	256
	Batchsize	$4/8/16$
	Activation	ReLU
	Meta Optimizer	SGD
	Optimizer	Adam
	Meta Learning rate	$1e-1$
	Align Learning rate	$1e-3$
	Classifier Learning rate	$1e-4$
	Iterations	800
	Loss	MSE
	Teacher forcing rate	0.5
	Number of evaluation tasks	16

between datapoints across languages. We chose a random $0.8-0.1-0.1$ split for train-val-test with respect to language for (79 languages, 9 languages, 10 languages), and the number of samples is 4395, 549, 549 for meta-train, meta-validation, and meta-test respectively. There is no overlap between the data used for source classification, target classification, and alignment tasks.

Hyperparameters: We show the hyperparameters used in Table 5.

9 Additional Results

Here we present some additional results, ablation studies, observations, and analysis on our approach.

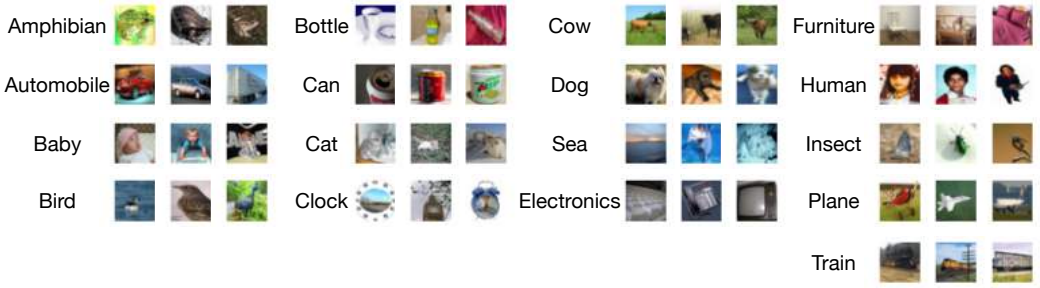

Figure 9: The 17 concepts shared across image and audio classification tasks that were used for weak alignment. Note that we only show the images - the audio spectrograms make up the second modality in each weak cluster.

Figure 10: On Yummly-28K dataset, CROMA leverages source text modality to make accurate few-shot predictions on target image modality despite only seeing $1 - 10$ labeled image examples.

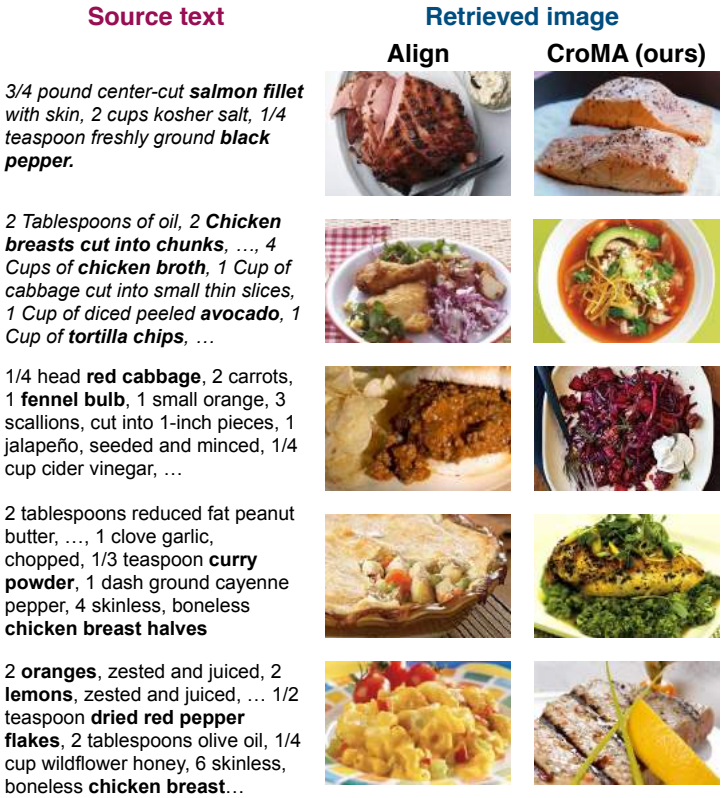

Figure 11: More samples of retrieved images given text recipes. CROMA performs few-shot retrieval of images more accurately than existing alignment approaches.

9.1 Text to Image

Extra results: We show some samples of image to recipe label predictions on low-resource image samples in Figure 10. Despite seeing just 5 labeled image samples, CROMA is able to quickly generalize and recognizes images from new recipes.

In Figure 11, we show more samples of retrieved data in the target given input in the source modality to help us understand which source modalities the model is basing its target predictions on. Our model yields better retrieval performance than the baselines, indicating that meta-alignment successfully aligns new concepts in low-resource target modalities.

9.2 Image to Audio

Extra results: We implement one more baseline derived as variations from existing work and adapted to our cross-modal generalization task. We adapt unsupervised meta-learning [26] which uses the aforementioned 17 weak clusters as prediction targets for the target modality during meta-training. This gives more discriminative training signal than the self-supervised reconstruction objective discussed in main text while still not explicitly using target modality labels during meta-training. We show these results in Table 6. While this baseline does do better than the reconstruction version of unsupervised meta-learning, it still underperforms as compared to CROMA.

9.3 Text to Speech

Extra results: We present some extra results by comparing with existing baselines in domain adaptation and transfer learning (see Section 10.2) in Table 7. We observe that none of them perform well on cross-modal generalization. Although domain confusion and alignment do improve upon standard encoder sharing, they still fall short of our approach. This serves to highlight the empirical differences between cross-modal generalization and domain adaptation. Therefore, we conclude that **1. separate encoders** and **2. explicit alignment** is important for cross-modal generalization and distinguishes it from domain adaptation.

10 Cross-modal Generalization vs Domain Adaptation

In this section we make both methodological and empirical comparisons with a related field of work, domain adaptation.

10.1 Methodological Differences

At a high-level, the core differences between cross-modal generalization and domain adaptation lies in the fact that domain adaptation assumes that both source and target data are from the same modality (e.g. image-only). As a result, these models are able to share encoders for both source and target domains [60]. This makes the alignment problem straightforward for this simplified version of the problem.

By sharing encoders, these domain adaptation methods do not directly model $p(x_s, x_t|m_s, m_t)$ for two different modalities, which does not provide the generalization guarantees we derived in Proposition 1. Without alignment, and domain adaptation is unlikely to work well since $p(x_s, x_t|m_s, m_t)$ is not modeled directly except on a few anchor points that some methods uses explicitly [68]. On the other hand, our approach explicitly models $p(x_s, x_t|m_s, m_t)$ using meta-alignment which in turn provides the guarantees in Proposition 1, thereby helping cross-modal generalization to low-resource modalities and tasks.

10.2 Empirical Differences

To further emphasize these methodological differences, we modify several classical domain adaptation methods for our task to verify that it is indeed necessary to use separate encoders and perform explicit alignment for cross-modal generalization. In particular, we implement the following baselines:

1. **Shared:** We share encoders for both modalities as much as much possible. The only non-shared parameter is a linear layer that maps data from the target modality's input dimension to the source so that all subsequent encoder layers can be shared. This reflects classical work in domain adaptation and transfer learning [29, 59].

Table 6: Performance on image to audio concept classification from CIFAR-10 and CIFAR-100 to ESC-50. CROMA is on par with the oracle few-shot audio baseline that has seen a thousand of labeled audio samples and outperforms existing unimodal and cross-modal baselines. #Audio (labeled) denotes the number of audio samples and labels used during meta-training.

TYPE	APPROACH	1-SHOT	5-SHOT	10-SHOT	#AUDIO (LABELED)
Uni	Unsup. pre-training [3, 12]	44.2 ± 0.8	72.3 ± 0.3	77.4 ± 1.7	0(0)
	Unsup. meta-learning [26] (reconstruct)	36.3 ± 1.8	67.3 ± 0.9	76.6 ± 2.1	920(0)
	Unsup. meta-learning [26] (weak labels)	45.6 ± 1.3	74.2 ± 0.3	83.7 ± 0.1	920(0)
Cross	Align + Classify [10, 24, 50, 59, 62]	45.3 ± 0.8	73.9 ± 2.1	78.8 ± 0.1	920(0)
	Align + Meta Classify [53]	47.2 ± 0.3	77.1 ± 0.7	80.4 ± 0.0	920(0)
	CROMA (ours)	$\mathbf{47.5 \pm 0.2}$	$\mathbf{85.9 \pm 0.7}$	$\mathbf{92.7 \pm 0.4}$	920(0)
Oracle	Within modality generalization [17, 45]	45.9 ± 0.2	89.3 ± 0.4	94.5 ± 0.3	920(920)

Table 7: Performance on text to speech generalization on the Wilderness dataset. We compare CROMA with some standard domain adaptation baselines and observe that none of them perform well on cross-modal generalization. Although domain confusion and alignment do improve upon standard encoder sharing, they still fall short of our approach. This serves to highlight the empirical differences between cross-modal generalization and domain adaptation.

TYPE	APPROACH	1-SHOT	5-SHOT	10-SHOT	#SPEECH (LABELED)
Domain Adaptation	Shared	55.6 ± 10.2	75.2 ± 8.4	81.9 ± 3.9	4395(0)
	Shared + Align [31]	59.7 ± 7.6	78.4 ± 6.2	84.3 ± 1.5	4395(0)
	Shared + Domain confusion [60]	59.5 ± 7.2	76.3 ± 9.4	83.9 ± 1.8	4395(0)
	Shared + Target labels [30]	57.3 ± 9.3	76.2 ± 8.4	84.0 ± 1.9	4395(4395)
Cross-modal	Align + Classify [10, 24, 50, 59, 62]	61.1 ± 6.0	74.8 ± 2.1	86.2 ± 0.7	4395(0)
	Align + Meta Classify [53]	65.6 ± 6.1	89.9 ± 1.5	93.0 ± 0.5	4395(0)
	CROMA (ours)	$\mathbf{67.9 \pm 6.6}$	$\mathbf{90.6 \pm 1.5}$	$\mathbf{93.2 \pm 0.2}$	4395(0)

2. **Shared + Align:** We share encoders for both modalities and further add our alignment loss (contrastive loss) on top of the encoded representations, in a manner similar to our meta-alignment model (a similar reference in the domain adaptation literature would be [31]).

3. **Shared + Domain confusion:** We share encoders for both modalities and further add a domain confusion loss on top of the encoded representations [60].

4. **Shared + Target labels:** Finally, we share encoders for both modalities and also use target modality labels during meta-training, in a manner similar to supervised domain adaptation [30].

Results: We present these results in Table 7 and observe that none of them perform well on cross-modal generalization. Although domain confusion and alignment do improve upon standard encoder sharing, they still fall short of our approach. Our method also outperforms the Shared + Target labels baseline which uses target modality labels to train the shared encoder during meta-training. This serves to highlight the empirical differences between cross-modal generalization and domain adaptation. Therefore, we conclude that **1. separate encoders** and **2. explicit alignment** is important for cross-modal generalization which distinguishes it from domain adaptation.

11 On Weak Alignment

This section discusses some mathematical guidelines on applying our method. Future theoretical work will be directed at formalizing the discussion in this section. For example, rigorous bounds on the minimizers can be derived when the models used are Lipschitz-continuous.

We focus on providing a understanding weak alignment before extending the analysis to cover the case of strong alignment. Let S denote the total number of weak-alignment sets, each with ρ^2 inner-set variance, and N_t be the number of target data points with supervision, then, clearly, a *tradeoff* in ρ^2 and $\frac{1}{N_t}$ exists: direct supervised learning results in a generalization error proportional to $\frac{1}{N_t}$, while weak supervision results in error proportional to ρ^2. Dividing N data points into S nearest neighbor sets, the resulting sets each have roughly N/S data points. If the original data points are drawn from a uniform distribution, then, each set will have variance proportional to $\frac{1}{S}$. Then, performing weak alignment is better than doing supervised learning if

$$\frac{c_s}{S} < \frac{c_t}{N_t} \tag{5}$$

for some architecture and task dependent constants $c_s,\ c_t$. This means that if the number of anchored sets is large or when the number of supervised data point is very small, then one should opt for using weak-alignment.

We can also rewrite this in terms of the number of data points for each set N_s we have. Since S is the number of anchor points, one expects that the error in alignment decreases as $\frac{1}{S}$. Let $N = SN_s$ denote the total number of data points. for some architecture and task dependent constant $c_s,\ c_t$. The above inequality is equivalent to

$$\frac{N_t}{S} = \frac{N_s N_t}{N} < c, \tag{6}$$

for some constant c. If we keep both the number of datapoints in each set and the supervised datapoints constant, then the trade-off depends only on N. If the number of total datapoints is large, one should use weak-alignment. What is the difference between learning with strong alignment and weak alignment? Intuitively, one would expect the generalization error to vanish when $N \to \infty$ for strong alignment, since the perfect one-to-one mapping between the target and the source can be discovered in this case. For weak alignment, however, one does not achieve vanishing generalization error in principle, since a fundamental uncertainty of order ρ^2 exists regarding the pairing relationship between different points within a given pair of anchored sets even if $N \to \infty$.

11.1 How to Choose S?

In the previous section, we assumed that the center for each set is known. However, it might come as a problem in practice if the sets are not given *a priori* and if one has to resort to clustering methods such as $k-$means for finding the desired sets and estimating their centers. In this case, one has fix N, but variable S and N_s. The error in alignment now depends on both S and N_s: (1) as S gets small, then the error, as discussed in the previous section, increases as $\frac{1}{S}$; (2) smaller N_s makes it harder for us to estimate the center of each set, and the by the law of large numbers, we can estimate the center at error of order $\frac{1}{N_s}$. This incurs an error of order

$$\frac{c_1}{S} + \frac{c_2}{N_s} = \frac{c_1}{S} + \frac{c_2 S}{N} > 0$$

for some constants $c_1,\ c_2$. One can take derivative to find the optimal S^* such that the error is minimized:

$$-\frac{c_1}{S^2} + \frac{c_2}{N} = 0 \to S^* = \sqrt{\frac{c_1 N}{c_2}}, \tag{7}$$

i.e. S^* should scale with $\sqrt{N}$.

11.2 Empirical Analysis

In this section, we verify our theoretical analysis on a controlled synthetic dataset. We generated synthetic data from 2 modalities: the source modality $D_1^{\text{sup}} = \{(x_1^i, y_1^i)\}_{i=1}^{n_1}$ and the target modality $D_2^{\text{sup}} = \{(x_2^i, y_2^i)\}_{i=1}^{n_2}$. The labels are generated via a noisy teacher model $y_m^i = u_m x_m^i + \epsilon_m^i$, where $x_m^i \in \mathbb{R}^d$, $u_m \in \mathbb{R}^d$, and $\epsilon_m^i \sim \mathcal{N}(0, \sigma^2)$ for $m \in \{1, 2\}$ [39]. We model cross-modal and task relationships through a full-rank transformation $x_1^i = W x_2^i$ and $u_1 x_1^i = u_2 W x_2^i$ respectively. In other words, we first sample points x_2 from the target modality from a chosen Gaussian distributions, and obtain points x_1 from the source modality via a cross-modal linear transformation W.

We consider the setting where we have a high-resource source task and a low-resource target task, so $n_1 \gg n_2$. One can train separate supervised models $f_m(x) = w_m x$ and measure the total generalization loss:

$$L = \sum_{m=1}^{2} \mathbb{E}_{x_m} \left[\left(f_m(x_m) - u_m x_m \right)^2 \right], \tag{8}$$

but this loss will be very high in the low-resource target task due to a very small number of labeled samples. Instead, cross-modal alignment learns the transformation W using pairs $D_{\text{unsup}} = \{(x_1^i, x_2^i)\}_{i=1}^{n_{\text{align}}}$ generated via $x_1^i = W x_2^i + \eta^i$ with noise $\eta^i \sim \mathcal{N}(0, \sigma_W^2)$. η^i models uncertainty in alignment pairs: $\sigma_W^2 \to 0$ represents strong alignment and large σ_W^2 represents weak alignment. By training a supervised model in the high-resource source modality together with learning cross-modal alignment, we are able to generalize to the low-resource target modality.

We empirically study this setup in Figure 12, where we set $d = 20, n_1 = 250, n_2 = 40$ and vary n_{align}. We make the following observations:

1. More alignment pairs help, but at most by the performance of the high-resource source task.
2. Quality of alignment matters: less noise σ_W^2 in alignment data gives better performance.
3. Even weak alignment is preferable to supervised learning with enough weakly paired data.

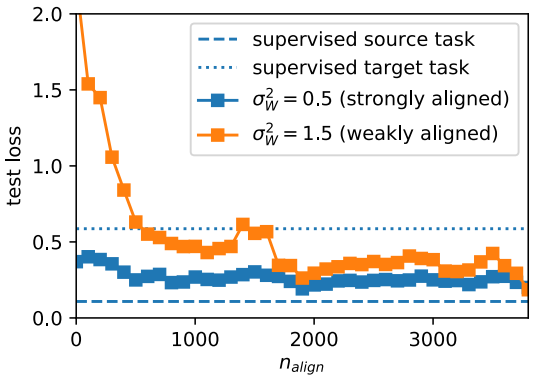

Figure 12: Supervision learning vs alignment for synthetic data: with more strongly or weakly aligned pairs (n_{align}), cross-modal alignment improves upon supervised learning in target tasks.

In fact, under this simplified setup, an analysis shows that training on the high-resource task has error $d\sigma^2/n_1$ while the low-resource task has error $d\sigma^2/n_2$. Estimating the alignment matrix with d^2 elements results in error $d^2\sigma_W^2/n_{\text{align}}$. Therefore, cross-modal alignment has error $\frac{d^2\sigma_W^2}{n_{\text{align}}} + \frac{d\sigma^2}{n_1}$, which should be preferred when $\frac{d\sigma_W^2}{n_{\text{align}}} + \frac{\sigma^2}{n_1} < \frac{\sigma^2}{n_2}$. This gives a simple rule-of-thumb for practitioners to choose between supervised learning and cross-modal learning.