Cross-Modal Discrete Representation Learning

Alexander H. Liu  SouYoung Jin  Cheng-I Jeff Lai  Andrew Rouditchenko
Aude Oliva  James Glass

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{alexhliu, souyoung, clai24, roudi, oliva, glass}@mit.edu

Abstract

Recent advances in representation learning have demonstrated an ability to represent information from different modalities such as video, text, and audio in a single high-level embedding vector. In this work we present a self-supervised learning framework that is able to learn a representation that captures finer levels of granularity across different modalities such as concepts or events represented by visual objects or spoken words. Our framework relies on a discretized embedding space created via vector quantization that is shared across different modalities. Beyond the shared embedding space, we propose a Cross-Modal Code Matching objective that forces the representations from different views (modalities) to have a similar distribution over the discrete embedding space such that cross-modal objects/actions localization can be performed without direct supervision. In our experiments we show that the proposed discretized multi-modal fine-grained representation (e.g., pixel/word/frame) can complement high-level summary representations (e.g., video/sentence/waveform) for improved performance on cross-modal retrieval tasks. We also observe that the discretized representation uses individual clusters to represent the same semantic concept across modalities.

1 Introduction

Toddlers acquire much of their knowledge through grounded learning – visual concepts can be acquired through language, and language acquisition emerges through visual interaction. Inspired by this type of grounded learning, a rich body of representation learning research [15, 28, 1, 31, 27] has been exploring the potential to learn from multi-modal data such as video-text, video-audio, and image-audio pairs. These works typically focus on learning a joint embedding space between different modalities, in which high-level summary representations are extracted as embedding vectors. These embedding vectors often represent entire video clips, spoken utterances, or sentences as single vectors, and can be useful on tasks such as cross-modal data retrieval, e.g., finding the most similar visual scene according to a spoken language description. The predominant approach to learning these embedding vectors is to use modality-independent encoders, and while this has been successful for downstream retrieval tasks, it makes it difficult to compare the activations of the encoders from different modalities. Further, the space of continuous embedding vectors is unbounded, which makes interpreting the learned representations challenging.

To this end, we propose to jointly learn high-level embedding vector representations with a fine-grained discrete embedding space that is shared across different modalities. The discrete embedding space enables model interpretability since there are a finite number of embedding vectors which are shared across modalities. Besides the shared embedding space, we propose a Cross-Modal Code
Matching (CMCM) objective that guides the embedding space to capture cross-modal correspondences of concepts, actions, and words. This not only improves downstream performance on retrieval, but also allows us to better interpret what the model recognized through cross-modal grounded learning.

To verify the effectiveness of our proposed learning framework, we conducted experiments in several cross-modal domains, including video-text, video-audio, and image-audio. We found consistent improvements over baseline models, verifying that the gain was not restricted to the particular choice of network architecture, input modalities, or dataset. We also demonstrate the interpretability of the fine-grained discrete representations by showing the cross-modal relations between the embedding vectors and semantic concepts appearing in the input modalities. Our approach also enables cross-modal concept localization without requiring any labels during training.

2 Methodology

Figure 1 provides an overview of the proposed framework. We begin by describing the two-branch cross-modal representation learning paradigm in Section 2.1 (the blue and yellow regions). Next, we introduce our shared discrete embedding space in Section 2.2 (the green region). Finally, in Section 2.3 and Figure 2, we introduce the Cross-Modal Code Matching objective which guides the model to learn semantically meaningful representations through the shared discrete embedding space.

2.1 Cross-Modal Learning Paradigm

Given a set of data $X = \{(x_i^A, x_i^B)\}_{i=1}^N$ of size $N$ where each instance $x_i$ is instantiated in different modalities $A$ and $B$ (e.g. video and its corresponding caption), the goal is to derive high-level representative vectors $(z_i^A, z_i^B)$ for each instance $(x_i^A, x_i^B)$ that capture the cross-modal relation measured by a choice of similarity function $S(\cdot, \cdot)$.

For a specific modality $M \in \{A, B\}$, a common first step is to encode raw data $x_i^M$ into a sequence of “fine-grained” latent features $H_i^M$ with a modality-specific neural network $f_{\text{fine}}^M$, i.e. $H_i^M = f_{\text{fine}}^M(x_i^M)$. The fine-grained representations $H_i^M$ can express different kinds of raw data, such as video, audio, or sentences, as a sequence of vectors $\{h_i^M_{1}, ..., h_i^M_{L}\}$ of length $L$. In the second step, a “high-level” representation $z_i^M$ can be derived by summarizing the fine-grained latent features $H_i^M$ with another encoding function $f_{\text{high}}^M$ that reduces the sequence into a single vector, i.e. $z_i^M = f_{\text{high}}^M(H_i^M)$.

For example, with modality $A$ being video, raw data $x_i^A$ can be treated as a sequence along time and space and encoded into fine-grained representations $H_i^A = \{h_i^A_{l}\}_{l=1}^L$ by choosing $f_{\text{fine}}^A$ to be a Residual Network [18]. For the second step, a natural choice for $f_{\text{high}}^A$ to derive the high-level representation $z_i^A$ would be a mean pooling function over the time and spatial axes (arranged along $l$).

With the sets of high-level representations $\{z_i^A\}_{i=1}^N$ and $\{z_j^B\}_{j=1}^N$ from different modalities, we can measure the cross-modal relation between any pair of representations $(z_i^A, z_j^B)$ with some similarity.
We trained the shared embedding space jointly with the rest of the framework by modifying the loss function \( \mathcal{L}_{\text{MMS}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{S(z_i^A, z_i^B) - M}}{e^{S(z_i^A, z_i^B) - M} + \sum_{j=1}^{N} I_{i \neq j} e^{S(z_i^A, z_j^B)}} \)\(^1\), where the margin \( M \) is a hyperparameter to encourage a higher similarity for positive pairs. The MMS loss \( \mathcal{L}_{\text{MMS}} \) can be seen as an application of the InfoNCE [33] loss with a margin.

The effectiveness of the described cross-modal learning paradigm has been shown by recent works that achieved state-of-the-art results on benchmark datasets in different cross-modal scenarios such as video-text [27], video-audio [31, 38], and image-text [36, 39].

### 2.2 Shared Discrete Embedding Space

While the high-level representations \((z_i^A, z_i^B)\) given by the cross-modal learning paradigm benefit end tasks such as data retrieval, the representations cannot be easily interpreted by humans. To obtain fine-grained representations that are more interpretable, we introduce a Vector Quantization [34] (VQ) mechanism after obtaining the \( H_i^M \) representations. Formally, with an auxiliary embedding table \( E = \{e_1, e_2, ..., e_V\} \) of size \( V \), which we refer to as the codebook, vector quantization is performed on each fine-grained representation \( h_{i,l}^M \in H_i^M \) of modality \( M \in \{A, B\} \) with

\[
\bar{h}_{i,l}^M = f^M(h_{i,l}^M) + \text{sg}(e_v - f^M(h_{i,l}^M)),
\]

where \( f^M \) is a modality specific projection network to project the input to the shared embedding space, \( v = \arg\min_{k \in V} \|\bar{h}_{i,l}^M - e_k\|_2 \), and \( \text{sg}(\cdot) \) is the stop-gradient operator proposed in straight-through gradient estimation [3] that treats the input as constant during backpropagation. In other words, each vector \( h_{i,l}^M \) will be replaced by its nearest neighbor \( e_v \), which we refer to as the codeword, in the codebook \( E \). The codebook is randomly initialized and updated with the exponential moving average [34] given the fine-grained representations (more details in Section A of the Appendix).

We trained the shared embedding space jointly with the rest of the framework by modifying the high-level representations \( z_i^M \) to include the discretized fine-grained representations as

\[
z_i^M = f_{\text{high}}(H_i^M) + f_{\text{code}}(\bar{H}_i^M),
\]

where \( f_{\text{high}} \) is, similar to \( f_{\text{code}} \), the encoding function for summarizing the sequence of quantized fine-grained representations (e.g., an average pooling function over \( l \)). Having such a discrete embedding space allows humans to better interpret the learned embeddings since they are shared across modalities and there are a finite number of them.

### 2.3 Cross-Modal Code Matching

Ideally, the codebook should be shared across different modalities since the quantization method is independent to the input modality. However, as we demonstrate in Section E of the Appendix, the model will learn to partition the codebook into modality-specific subspaces due to the significant difference between fine-grained representations from different modalities. To learn a shared embedding space that is invariant to input modality, we propose the Cross-Modal Code Matching objective which encourages the model to focus more on the semantic aspect of the input, as illustrated in Figure 2.

For each vector \( h_{i,l}^M \) in the fine-grained representation sequence \( H_i^M \) encoded from an instance \( z_i^M \) of modality \( M \), we first define the probability of \( h_{i,l}^M \) belonging to the codeword \( e_v \) as the Softmin function of their Euclidean distance, that is

\[
P(e_v|h_{i,l}^M) = \frac{\exp(-\|f^M(h_{i,l}^M) - e_v\|_2)}{\sum_{k \in V} \exp(-\|f^M(h_{i,l}^M) - e_k\|_2)}.
\]

---

\(^1\)While we used dot product throughout this work, we also found euclidean distance works well in practice.
Figure 2: Our proposed Cross-Modal Code Matching objective (described in Section 2.3), which encourages the model to use similar codewords for matching cross-modal pairs.

Note that this definition assigns higher a probability to codewords that are closer to the fine-grained representation, where the closest codeword is used to perform vector quantization. We can then define the sequence-level probability distribution over the codebook as the average of the fine-grained distribution, that is

\[ P(e_v | H^M) = \frac{1}{L} \sum_i P(e_v | h^M_{i,j}). \]  

(5)

This distribution is essentially the normalized frequency of codeword usage for a given sequence of fine-grained representations. Next, for a pair of cross-modal data \((x^A_i, x^B_j)\), we define their code similarity as the negative symmetric cross entropy of probability distribution over the codebook

\[ S_{\text{code}}(x^A_i, x^B_j) = \sum_v P(e_v | H^A_i) \log P(e_v | H^B_j) + \sum_v P(e_v | H^B_j) \log P(e_v | H^A_i), \]  

(6)

Finally, we propose the Cross-Modal Code Matching (CMCM) objective using code similarity as

\[ \mathcal{L}_{\text{CMCM}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{S_{\text{code}}(x^A_i, x^B_j)}}{e^{S_{\text{code}}(x^A_i, x^B_j)} + \sum_{j \neq i} e^{S_{\text{code}}(x^A_i, x^B_j)}}. \]  

(7)

Intuitively, the proposed objective encourages the model to represent the input \((x^A_i, x^B_j)\) with similar codewords for positive pairs \((i = j)\) and non-matching codewords for negative pairs \((i \neq j)\). As a consequence, each codeword is expected to be a modality invariant representation of a more fine-grained concept, action, or word that can be discovered from cross-modal data. For example, a codeword could correspond to both the visual scene of a man juggling, and also the spoken word “juggling,” as we demonstrate in our experimental results in Table 2 and Figure 4.

The full objective of our proposed cross-modal representation learning framework is the combination of objectives at different levels

\[ \mathcal{L} = \mathcal{L}_{\text{MMS}} + \alpha \mathcal{L}_{\text{CMCM}}, \]  

(8)

where \(\alpha\) controls the weight between the two terms. Empirically, we found \(\alpha = 0.1\) worked well across different settings.

3 Related work

Methods fitting into the cross-modal learning paradigm. As described in Section 2.1, many of the existing methods for cross-modal learning fit into the paradigm where encoders are modality-independent. This paradigm has been shown to be effective by achieving state-of-the-art retrieval performance on benchmark datasets with the modality pairs that we considered in this work: video-text [2, 27], video-audio [31, 38], and image-audio [15, 14]. While these prior works relied on different pre-training datasets, model architectures, and objective functions, they all leverage modality-independent encoders. One of the most important features of this paradigm is the fixed inference time for retrieval. Since the encoders are modality-independent, embedding vectors for samples in a given modality can be computed without using any samples from the other modality. Thus retrieval only involves computing the dot product between embedding vectors from two different modalities. As a consequence, these models are more flexible for large-scale retrieval, and the embedding vectors from each modality can be used independently for other downstream tasks.
Other cross-modal learning frameworks. In contrast to the aforementioned works, some methods leverage cross-modal relations within the encoders instead of using modality-independent encoders. This has been done with both cross-modal encoders [21, 27] and cross-modal attention mechanisms [29, 26, 25, 10]. However, the cross-modal interactions increase the complexity for retrieval since every instance of a specific modality must be used as input with every instance of another modality to obtain the embedding vectors. With m and n samples in the modalities respectively, this increases the complexity from the modality-independent approach from \( \mathcal{O}(m + n) \) to \( \mathcal{O}(mn) \). Further, it also makes analysis of the embedding vectors from any individual modality challenging and inhibits single-modality downstream tasks. Our proposed framework builds on the modality-independent approach to enable light-weight retrieval, but it also enables cross-modal interaction through our proposed codebook and Cross-Modal Code Matching objective.

Uncovering semantic-level correspondences. Image-audio models have been shown to discover spoken words and visual objects without supervision through retrieval tasks [41, 12, 17, 20], and the audio embedding vectors have been shown to cluster into word-like speech units [13, 43, 14]. Some work has studied the ability of video-audio models to relate spoken words to visual objects and actions in videos [4, 38]. However, none of these models incorporated a shared embedding space that enabled modality-invariant representations. VQ units have been used in the audio encoder of an image-audio model [14], which allowed it to capture the hierarchical structure of spoken language. While our proposed framework is similar in that it also discretizes the audio sequence with VQ units, our work differs significantly by capturing the cross-modal interactions between visual and audio inputs in the shared embedding space rather than solely capturing the tree structure of speech. Further, besides image-audio data, our proposed framework can handle video-audio and video-text data. Finally, modality invariant audio-visual representations have been explored using variational autoencoders [22, 47], while we propose modality invariant representations for different cross-modal domains using a shared discrete embedding space.

4 Experiments

4.1 Setup

To demonstrate the generalizability of the proposed method, we tested our framework on different cross-modal datasets and baseline models that fit into the cross-modal learning paradigm. All setups are listed below and summarized in Table 4 of the Appendix. For training the proposed model, we randomly initialized all the modules related to the discrete shared embedding space and trained them jointly with the rest of the framework (see Figure 1). Unless otherwise specified, (1) we “warm-started” our proposed framework by initializing it with the modality-specific encoders (namely, \( f_{\text{low}}^M \) and \( f_{\text{high}}^M \)) from the baseline models; (2) both the projection network \( f^M \) and the encoder network \( f_{\text{code}} \) are single linear layers; (3) the codebook size is set to 1024. Please refer to Section B in the Appendix for more implementation details and computational costs.

Video-Text: MSR-VTT [44] contains 10k video clips with length varying from 10 to 32 seconds. While each video is provided with 20 related captions without supervision through retrieval tasks [41, 12, 17, 20], and the audio embedding vectors have been shown to cluster into word-like speech units [13, 43, 14]. Some work has studied the ability of video-audio models to relate spoken words to visual objects and actions in videos [4, 38]. However, none of these models incorporated a shared embedding space that enabled modality-invariant representations. VQ units have been used in the audio encoder of an image-audio model [14], which allowed it to capture the hierarchical structure of spoken language. While our proposed framework is similar in that it also discretizes the audio sequence with VQ units, our work differs significantly by capturing the cross-modal interactions between visual and audio inputs in the shared embedding space rather than solely capturing the tree structure of speech. Further, besides image-audio data, our proposed framework can handle video-audio and video-text data. Finally, modality invariant audio-visual representations have been explored using variational autoencoders [22, 47], while we propose modality invariant representations for different cross-modal domains using a shared discrete embedding space.

Video-Audio: S-MIT [31] contains over 500k pairs of 3-second video and corresponding spoken audio captions averaging 8 seconds. We followed the official protocol to train on the training set of 500k pairs, use the validation set of 10k pairs for development and analysis, and report the retrieval result on a 1k search space over 5 runs randomly sampled from a held-out test set. We selected the same baseline model used on the dataset [31], which contains a visual encoder composed of a
### Table 1: Cross-Modal retrieval results on MSR-VTT, S-MiT, and Places.

| Modality A-B / Dataset | Visual Retrieval (B → A) | Language Retrieval (A → B) |
|------------------------|--------------------------|---------------------------|
|                        | R@1 ↑ | R@5 ↑ | R@10 ↑ | MnR ↓ | R@1 ↑ | R@5 ↑ | R@10 ↑ | MnR ↓ |
| Video-Text / MSR-VTT [44] |        |        |        |        |        |        |        |        |
| Frozen-in-Time [2]      | 31.0  | 59.5   | 70.5   | -      | -      | -      | -      | -      |
| CLIP4Clip-meanP [27]    | 43.1  | 70.4   | 80.8   | 16.2   | -      | -      | -      | -      |
| CLIP4Clip-tightT [27]   | 40.2  | 71.5   | 80.5   | 13.4   | -      | -      | -      | -      |
| Our Baseline†           | 42.6  | 71.2   | 80.8   | 15.5   | 43.0   | 70.9   | 80.9   | 12.5   |
| Proposed                | 43.4  | 72.3   | 81.2   | 14.8   | 42.5   | 71.2   | 81.1   | 12.0   |
| Video-Audio / S-MiT [31] |        |        |        |        |        |        |        |        |
| S-MiT [31]              | 32.1  | 58.9   | 68.6   | -      | 32.3   | 57.9   | 68.1   | -      |
| Our Baseline†           | 30.2  | 57.3   | 68.5   | 41.9   | 29.7   | 57.2   | 68.7   | 28.5   |
| Proposed                | 34.3  | 61.3   | 72.0   | 33.5   | 34.0   | 61.6   | 71.7   | 22.5   |
| Image-Audio / Places [17] |        |        |        |        |        |        |        |        |
| ResDAVEnet [16]†        | 30.9  | 63.6   | 74.2   | 20.2   | 26.4   | 58.5   | 71.2   | 21.6   |
| ResDAVEnet-VQ [14]*     | 34.9  | 70.2   | 79.4   | 15.0   | 32.7   | 65.6   | 77.0   | 18.0   |
| Our Baseline†           | 43.8  | 74.1   | 82.4   | 15.8   | 40.4   | 73.3   | 82.5   | 10.9   |
| Proposed                | 46.5  | 77.4   | 85.8   | 13.7   | 45.4   | 77.7   | 85.9   | 8.9    |

† Existing model reproduced with L_{mmse} for fair comparison, see Table 4 in the Appendix for more detail.
* Results obtained by running the official code and pre-trained models, see Appendix for more details.

ResNet-152 pre-trained on ImageNet [7] and TSM ResNet-50 [23] pre-trained on M-MiT [32]. The audio encoder is a randomly initialized 1D-ResNet [16] designed specifically for spectrograms. The shared embedding space has the dimension of 4096, matching the encoders in the baseline model.

**Image-Audio: Places [17]** contains over 400k pairs of images from the Places 205 dataset [46] and corresponding spoken audio captions averaging 10 seconds. We followed the previous works [15, 16, 14] to use the training set of 400k pairs and report results on the validation set of 1k pairs. We select ResDAVEnet [16] as the baseline model where the visual encoder is a ResNet-50 pre-trained on ImageNet [7] and the audio encoder is a randomly initialized 1D-ResNet [16] designed specifically for spectrograms. The shared embedding space has the dimension of 1024.

### 4.2 Cross-Modal Retrieval

Data retrieval is one of the most common evaluations for cross-modal representation learning. For example, in video retrieval with input query text, videos in the search space will be ranked by the similarity between the representation of each video and the query. We report the standard retrieval metrics recall at rank K (R@K) and median rank (MdR) in Table 1. We show the performance on both visual retrieval, where input language queries are used to retrieve videos or images, and language retrieval, where input visual queries are used to retrieve spoken or text captions.

**Video-Text Retrieval.** On the benchmark MSR-VTT dataset, we compared our proposed method against recent works achieving state-of-the-art [2, 24, 27] and provide a full comparison against more prior work [26, 38, 10, 35, 9, 5] in Section C of the Appendix. Frozen-in-Time [2] and CLIP4Clip [27] are similar methods that employ a Visual Transformer [8] to encode video as sequence of images. The key differences between them is the choice of summarizing function (i.e. \( f_M \)) for video and the pre-training procedure. We also note that the CLIP4Clip with tight transformer encoder [27] (CLIP4Clip-tightT) relied on cross-modal reference via self-attention encoders to derive representations, which has a higher time complexity as mentioned in Section 3. With the shared codebook and Cross-Modal Code Matching objection, our proposed framework also enables cross-modal reference and gives an improvement over the baseline model without increasing the time complexity.

**Video-Audio Retrieval.** Video-Audio retrieval on S-MiT [31] is a challenging task since videos are paired with raw speech audio, which is untranscribed, unsegmented, and can contain background noise and speaker variation. However, our proposed framework that leverages cross-modal connections
Figure 3: Conditional probability matrix illustrating $P(\text{action}|\text{codeword})$ on the S-MiT development set. Y-axis is action label, showing only the top 20 most frequent labels for simplicity. X-axis is the indices of the top 100 most frequent codewords.

between visual actions and spoken words is able to improve the baseline model by a margin. We further analyze our framework’s ability to relate visual actions and spoken words in Section 4.3.

**Image-Audio Retrieval.** We compare our proposed method against the recent models [16, 14] achieving state-of-the-art and provide a full comparison to previous methods [15, 13, 17] in Section D of the Appendix. Comparing the baseline model, ResDAVEnet [16], and the current state-of-the-art ResDAVEnet-VQ [14], the latter model introduces VQ units into the audio encoder, allowing it to model the hierarchical structure of speech and achieve better retrieval results. With our framework, we introduce our shared VQ embedding space into the ResDAVEnet model to capture cross-modal interactions. This improves the performance over both ResDAVEnet and ResDAVEnet-VQ.

Overall, our proposed method enables consistent improvements regardless of the data modalities and baseline architectures, demonstrating its effectiveness and generalizability.

### 4.3 Discrete Representation Analysis

One of the important motivations of introducing the discrete cross-modal embedding space is better model interpretability. In this section, we take a closer look into the codewords learned through our proposed framework. For the evaluation, we chose the video-audio setup on S-MiT [31]. We used video-audio pairs from the development set, where each pair is labeled with an action out of 332 categories. Note that we only used labels for analysis, labels are never used for training.

**Conditional Probability of Action Labels Given Codeword.** First, we compute the conditional probability distributions of action labels given the codewords over the video inputs. Each video input is fixed-length and represented by 27 codewords (3 frames each represented by 3×3 codewords), and we labeled all these codewords with the video’s action label. By accumulating codeword labels through the whole development set, we can compute the conditional probability of each action given any codeword, i.e. $P(\text{action}|\text{codeword})$. Results are visualized in the upper part of Figure 3. Similarly, we computed the conditional probabilities based on the audio input where each utterance is represented by up to 32 codewords depending on the utterance length. We selected the most frequent codewords used by the video inputs and plot the conditional probabilities based on the audio input in the lower part of Figure 3. We can observe that both matrices have similar patterns, i.e., when a codeword is activated, there is a high chance of a specific action appearing in the input regardless if it is video or audio. This suggests that our model is able to learn cross-modal representations for actions grounded by either visual or spoken language input. The codewords are not only modality invariant, but more importantly, they also capture the semantic relations of the labels, e.g., codewords with the highest chance to represent “autographing” typically have the second highest chance of representing “signing”; codewords for “surfing” are less likely to represent other actions as all of them are very different from “surfing”. We also note that without the Cross-Modal Code Matching objective,
Table 2: Correspondence between codewords, visual actions, and spoken words. Ranking is based on the precision (Prc.) of the top hypothesis of the visual action label. Occurrence (Occ.) indicates the number of times the codeword was activated throughout the development set. Around 750 codewords were activated on the development set. An extended table is available in Section F of the Appendix.

| Rank | Code | Occ. | Visual Action | Spoken word | Visual Action | Spoken word |
|------|------|------|---------------|-------------|---------------|-------------|
|      |      |      | Top Hypothesis | Second Hypothesis | Top Hypothesis | Second Hypothesis |
|      |      |      | Label | Prc. | Label | Prc. | Word | F1 | Word | F1 |
| 1    | 201  | 147  | juggling | 97.5 | kicking | 1.2 | juggling | 36.7 | juggles | 8.3 |
| 2    | 349  | 112  | flossing | 96.0 | licking | 0.7 | floss | 15.8 | flossing | 14.0 |
| 3    | 145  | 49   | surfing   | 95.6 | snowing | 2.9 | surfboard | 23.7 | waves | 7.3 |
| 4    | 29   | 64   | tattooing | 94.6 | injecting | 2.2 | tattoo | 15.8 | tattooed | 4.2 |
| 5    | 233  | 25   | ironing   | 93.8 | hammering | 6.2 | ironing | 20.5 | iron | 4.7 |
|      |      |      |          |      |          |      |      |      |      |      |
| 32   | 500  | 89   | dialing   | 60.0 | texting | 10.0 | dialing | 13.8 | phone | 9.8 |
| 33   | 536  | 28   | cheering  | 60.0 | shouting | 10.0 | cheerleaders | 26.8 | cheering | 10.3 |
| 34   | 50   | 203  | rafting   | 58.6 | paddling | 25.7 | rafting | 16.7 | raft | 8.5 |
| 35   | 664  | 78   | dunking   | 58.0 | leaping | 9.1 | basketball | 11.0 | dunking | 5.2 |
|      |      |      |          |      |          |      |      |      |      |      |
| 742  | 733  | 188  | discussing | 6.5 | applauding | 4.6 | men | 7.3 | two | 6.4 |
| 743  | 542  | 58   | baking    | 6.5 | peeling | 5.2 | cupcake | 9.2 | peanut | 6.2 |

Figure 4: Codeword cross-modal localization. Input regions that are encoded by the codeword (selected from Table 2) are highlighted in red.

semantically related video and audio inputs no longer use the same codewords, which we illustrate in Section E of the Appendix.

Cross-Modal Correspondences. Next, we analyze the connections captured by the codewords between action labels and spoken words. With the same label accumulation method described previously, we compute the precision of action prediction with codewords (i.e., code-action co-occurrence). For the audio, we used word-level transcriptions (from Google's speech-to-text API) to assign a spoken word to each codeword when it is activated by the input utterance. This results in a hypothesis set including around 7k words for each codeword, and we listed the top 2 hypotheses for each codeword with the highest F1 score (instead of precision to avoid domination of high-frequency words). Results are listed in Table 2. For the codewords that have the highest precision on predicting the action label, we found the top hypotheses for spoken words are often the action label itself. E.g., the codeword (rank 1st) for the visual action “juggling” maps to the spoken word “juggling” perfectly. As precision on visual action prediction decreases, we observed fewer perfect mappings, but the spoken word hypotheses remained semantically related to the visual action hypotheses. E.g., the codeword (rank 35th) for the visual action “dunking” with lower precision now maps to the spoken word “basketball.” Surprisingly, even the codewords with the lowest precision capture relationships between visual actions and spoken words to some extent. E.g., codeword (rank 743rd) that is most related to the action “baking” has the top and second word hypotheses “cupcake” and “peanut.”

Codeword Localization. Finally, to visualize the relation between codewords and the input data, we localize the segments of both the video and audio input that are assigned to certain codewords.
Table 3: Ablation study performed on Places [17], scores are averaged over audio and image retrieval.

| Method | Averaged 2-way Retrieval | R@1 ↑ | R@5 ↑ | R@10 ↑ | MnR ↓ |
|--------|---------------------------|------|-------|--------|-------|
| (a) Proposed | 46.0 | 77.6 | 85.9 | 11.3 |
| (b) codebook size = 512 | 46.2 | 77.4 | 85.2 | 11.5 |
| (c) codebook size = 2048 | 46.1 | 76.6 | 84.7 | 12.1 |
| (d) $\alpha = 1.0$ | 45.6 | 76.6 | 85.5 | 11.6 |
| (e) $\alpha = 0.0$ (w/o Cross-Modal Code Matching) | 45.2 | 75.5 | 84.2 | 12.8 |
| (f) w/o VQ & w/o Cross-Modal Code Matching | 45.7 | 75.9 | 84.7 | 12.6 |
| (g) w/o warm-start | 41.6 | 73.4 | 82.5 | 16.0 |
| (h) w/o continuous representation ($f_{\text{high}}^{M}(H_{M}^{f})$) | 29.0 | 63.0 | 74.7 | 19.4 |
| (i) Our Baseline | 42.1 | 73.7 | 82.5 | 13.4 |

This is possible because quantization in our shared embedding space is done at the fine-grained level, so that the time and spatial axes are preserved. Examples are shown in Figure 4, where the regions assigned to the given code are highlighted. Interestingly, we see the codewords being aligned to both the visual actions and the corresponding spoken words. This supports our claim of having a more interpretable representation at the fine-grained level.

4.4 Ablation Study

To justify our framework design and choice of hyperparameters, we conducted an ablation study on the image-audio setting and report the results in Table 3.

Impact of the shared embedding space. For the codebook size, 1024 codewords worked well across different datasets. Halving and doubling the number of codewords (row(b) & (c)) both decreased the performance slightly. For the weight $\alpha$ of the proposed Cross-Modal Code Matching objective, we found that values in the range $(0, 1]$ generally work while 0.1 works the best (row(a) v.s. row(d)). Removing the proposed Cross-Modal Code Matching objective (setting $\alpha = 0.0$, row(e)), however, hurts the performance. Furthermore, without the objective, the codebook no longer captures cross-modal correspondences, as illustrated in Section E of the Appendix. We also observed that disabling the VQ layer together with the Cross-Modal Code Matching objective slightly recovers performance (row(f) v.s. row(e)). All of these observations serve as evidence that the proposed discrete embedding space is most beneficial to the retrieval task with the guidance from the Cross-Modal Code Matching objective.

Importance of baseline models in the cross-modal learning paradigm. As mentioned in Section 4.1, the discrete shared embedding space is learned with “warm-starting” from a baseline model. We note that warm-starting is important for getting more refined representations that yield better retrieval results (row(a) v.s. row(g)). Without warm-starting, our framework can only perform similar to the baseline (row(g) v.s. row(i)). This finding aligns with previous work [14] that used VQ layers in the audio encoder and used warm-starting to learn acoustic units. Moreover, removing the continuous representations (row(h)) originally used in the cross-modal learning paradigm and using only the codeword representations significantly decreases performance. This exposes the trade-off between interpretability and end-task performance by imposing a discrete embedding space. Hence, we choose to integrate both discrete and continuous embedding space for retrieval as in Eq. 3.

5 Conclusion

In this paper, we proposed a framework for cross-modal representation learning with a discrete embedding space that is shared amongst different modalities and enables model interpretability. We also propose a Cross-Modal Code Matching objective that encourages models to represent cross-model semantic concepts in the embedding space. Combining our discrete embedding space and objective with existing cross-modal representation learning models improves retrieval performance on video-text, video-audio, and image-audio datasets. We also analyze the shared embedding space and find that semantically related video and audio inputs tend to use the same codewords.

Limitations. As described in Section 2.1, the present work relies on the existing cross-modal learning paradigm with modality-independent encoders. Although our proposed method is shown
to be effective, further work is required to generalize our method to other cross-modal learning frameworks which are more computationally complex. In addition, recent work [1, 38] demonstrates the benefits of learning from three or more modalities. While the proposed method can theoretically be extended to more modalities, further work is required to handle practical bottlenecks such as the growth in time complexity for the Cross-Modal Code Matching objective with respect to the number of modalities.

Potential Negative Societal Impacts. Our proposed method is self-supervised and does not use any labels during training. It will therefore learn any biases present in the data. However, we expect that our discrete embedding space will help improve interpretability and could help humans discover the biases present in training data before deploying models to any real-world applications.

Acknowledgments and Disclosure of Funding

This research was supported by the MIT-IBM Watson AI Lab. We thank Dave Harwath for helpful discussions.

References

[1] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, and A. Zisserman. Self-supervised multimodal versatile networks. arXiv preprint arXiv:2006.16228, 2020.
[2] M. Bain, A. Nagrani, G. Varol, and A. Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. arXiv preprint arXiv:2104.00650, 2021.
[3] Y. Bengio, N. Léonard, and A. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432, 2013.
[4] A. W. Boggust, K. Audhkhasi, D. Joshi, D. Harwath, S. Thomas, R. S. Feris, D. Gutfreund, Y. Zhang, A. Torralba, M. Picheny, et al. Grounding spoken words in unlabeled video. In CVPR Workshops, pages 29–32, 2019.
[5] I. Croitoru, S.-V. Bogolin, Y. Liu, S. Albanie, M. Leordeanu, H. Jin, and A. Zisserman. Teachtext: Crossmodal generalized distillation for text-video retrieval. arXiv preprint arXiv:2104.08271, 2021.
[6] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier. Language modeling with gated convolutional networks. In International conference on machine learning, pages 933–941. PMLR, 2013.
[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.
[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
[9] M. Dzabraev, M. Kalashnikov, S. Komkov, and A. Petiushko. Mdmm: Multidomain multimodal transformer for video retrieval. arXiv preprint arXiv:2103.10699, 2021.
[10] V. Gabeur, C. Sun, K. Alahari, and C. Schmid. Multi-modal transformer for video retrieval. In European Conference on Computer Vision (ECCV), volume 5. Springer, 2020.
[11] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Y. W. Teh and M. Titterington, editors, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, volume 9 of Proceedings of Machine Learning Research, pages 297–304, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.
[12] D. Harwath and J. Glass. Deep multimodal semantic embeddings for speech and images. In 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pages 237–244. IEEE, 2015.
[13] D. Harwath and J. R. Glass. Learning word-like units from joint audio-visual analysis. arXiv preprint arXiv:1701.07481, 2017.
[14] D. Harwath, W.-N. Hsu, and J. Glass. Learning hierarchical discrete linguistic units from visually-grounded speech. In International Conference on Learning Representations, 2020.
[15] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass. Jointly discovering visual objects and spoken words from raw sensory input. In Proceedings of the European conference on computer vision (ECCV), pages 649–665, 2018.
[16] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass. Jointly discovering visual objects and spoken words from raw sensory input. IJCV, 2020.
[17] D. Harwath, A. Torralba, and J. R. Glass. Unsupervised learning of spoken language with visual context. In Neural Information Processing Systems, 2017.
[44] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5288–5296, 2016.

[45] Y. Yu, J. Kim, and G. Kim. A joint sequence fusion model for video question answering and retrieval. In Proceedings of the European Conference on Computer Vision (ECCV), pages 471–487, 2018.

[46] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In NeurIPS, 2014.

[47] Y. Zhu, Y. Wu, H. Latapie, Y. Yang, and Y. Yan. Learning audio-visual correlations from variational cross-modal generation. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4300–4304. IEEE, 2021.
Appendix

A Codebook Update Policy

The codebook with \(d\)-dimensional codewords is initialized with

\[
N_v^{(0)} = 1 \\
m_v^{(0)} \sim \mathcal{N}(0, 1) \\
e_v^{(0)} = m_v^{(0)},
\]

and updated with each codeword \(e_v\) being the exponential moving average (EMA) of all the fine-grained representations \(H = \{f^M(M_{i,l}) \mid \overline{h}_{i,l} = e_v\}\) that was replaced by \(e_v\) for every training step \(t\):

\[
N_v^{(t)} \leftarrow \gamma N_v^{(t-1)} + (1 - \gamma) |H| \\
m_v^{(t)} \leftarrow \gamma m_v^{(t-1)} + (1 - \gamma) \sum h \\
e_v^{(t)} \leftarrow \frac{m_v^{(t)}}{N_v^{(t)}},
\]

where the decay factor \(\gamma\) is set to 0.99 throughout this work. To improve the overall usage of the codebook, the input fine-grained representations are modality-wise batch normalized. In addition, codewords that are not activated (i.e. \(|H| = 0\)) for 100 consecutive steps are re-initialized during codebook update. The reset value is randomly chosen from activated codewords.

B Implementation Details

For each dataset and modality pair considered in this work, we selected baseline models that follow the cross-modal learning paradigm (as described in Section 2.1). Baseline models with different fine-grained and high-level encoders (\(f_{\text{fine}}^M\) and \(f_{\text{high}}^M\)) are summarized in Table 4. The links to the official implementation of these baseline models are also provided in the table. For a fair comparison, we retrained the models with the \(L_{\text{MMS}}\) (margin set to 1e-3) as our baseline models.

| Setup       | Modality | Encoders from baseline model |
|-------------|----------|------------------------------|
| Dataset     | Baseline model |                   |
| MSR-VTT [44] | video    | Vision Transformer\(^3\)\[^8\] | Avg. Pooling + Linear |
|             | - CLIP4Clip\(^1\) [27] | Transformer\(^4\)\[^42, 37\] [EOT] token + Linear |
| S-MiT [31]   | text     | ResNet-152\(^5\) [18] + TSM\(^6\) [23] | Max Pooling + GLU [6] |
|             | - AVLnet\(^2\) [38] | audio Spectrogram+1D-ResNet [16] | Avg. Pooling + GLU [6] |
| Places [17]  | image    | ResNet-50\(^7\) [18] | Avg. Pooling + GLU [6] |
|             | - ResDA VEnet\(^3\) [16] | audio Spectrogram+1D-ResNet [16] | Avg. Pooling + GLU [6] |

\(^1\) [https://github.com/ArrowLuo/CLIP4Clip]
\(^2\) [https://github.com/roudimit/AVLnet (under BSD license)]
\(^3\) [https://github.com/wnhsu/ResDAVEnet-VQ (under BSD license)]
\(^4\) Initialized from CLIP model pretrained on WebImageText dataset [36].
\(^5\) Pretrained on ImageNet [7].
\(^6\) Pretrained on Multi-MiT [32].

MSR-VTT. For our baseline model, we did not reproduce CLIP4Clip’s post-pretraining stage, which trained CLIP4Clip on the subset of HowTo100M [30] before adapting to MSR-VTT, since this stage is not necessary for the best results on MSR-VTT and the subset is not released. We used all of the
hyper-parameters of the official implementation except the batch size is reduced from 128 to 64 to meet our hardware restriction. To train the shared discrete embedding space, we warm-started from the baseline model with a learning rate of 1e-5. Each video is encoded into 8 codewords ($2 \times 2 \times 2$ for time, height, width) and each subword unit in the sentence is encoded into 1 codeword. The baseline model is trained for 12 hours on 8 2080Ti GPUs; and it takes an additional 6 hours to train the proposed framework.

**S-MiT.** The input audio feature is a 40 dimensional mel-spectrogram with a window size of 25 ms and a hop size of 10 ms. The baseline is trained with a batch size of 2048 and a learning rate of 1e-3. To train the shared discrete embedding space, we warm-started from the baseline model with a learning rate of 1e-4. Each video is encoded into 27 codewords ($3 \times 3 \times 3$ for time, height, width) and every 16 consecutive frames from the spectrogram is encoded into 1 codeword. The baseline model is trained for 4 hours on 4 V100 GPUs; and it takes an additional 1 hour to train the proposed framework. For both baseline model and our proposed model, we followed the previous work [31] to perform a second round training with a learning rate of 1e-5 and a batch size of 128. The second round training fine-tunes the TSM video encoder (which is frozen in the first round training) on S-MiT jointly with the rest of the components, which takes 2 days on 8 Titan RTX GPUs.

**Places.** The input audio feature is a 40 dimensional mel-spectrogram with a window size of 25 ms and a hop size of 10 ms. The baseline is trained with a batch size of 256 and a learning rate of 1e-3. To train the shared discrete embedding space, we warm-started from the baseline model with a learning rate of 1e-4. Each image is encoded into 49 codewords ($7 \times 7$ for height, width) and every 16 consecutive frames from the spectrogram is encoded into 1 codeword. The baseline model is trained for 36 hours on 1 V100 GPU; and it takes an additional 4 hours to train the proposed framework.

### C MSR-VTT Video Retrieval Full Comparison

| Method                                                                 | Video Retrieval (Text $\rightarrow$ Video) | R@1 ↑ | R@5 ↑ | R@10 ↑ | MnR ↓ |
|------------------------------------------------------------------------|--------------------------------------------|--------|--------|--------|-------|
| Collaborative Experts [26]                                             |                                            | 20.9   | 48.8   | 62.4   | 28.2  |
| Multi-Modal Transformer [10]                                           |                                            | 26.6   | 57.1   | 69.6   | 24.0  |
| Support-Set Bottlenecks [35]                                           |                                            | 30.1   | 58.5   | 69.3   | -     |
| Multidomain Multimodal Transformer [9]                                 |                                            | 38.9   | 69.0   | 79.7   | 16.5  |
| Frozen-in-Time [2]                                                     |                                            | 31.0   | 59.5   | 70.5   | -     |
| Hierarchical Transformer with Momentum Contrast [24]                   |                                            | 30.7   | 60.9   | 73.2   | -     |
| TeachText [5]                                                          |                                            | 29.6   | 61.6   | 74.2   | -     |
| CLIP4Clip-meanPooling [27]                                             |                                            | 43.1   | 70.4   | 80.8   | 16.2  |
| CLIP4Clip-seqLSTM [27]                                                 |                                            | 42.5   | 70.8   | 80.7   | 16.7  |
| CLIP4Clip-seqTransformer [27]                                          |                                            | 44.5   | 71.4   | 81.6   | 15.3  |
| CLIP4Clip-tightTransformer [27]                                        |                                            | 40.2   | 71.5   | 80.5   | 13.4  |
| Our Baseline (based on CLIP4Clip-meanPooling)                         |                                            | 42.6   | 71.2   | 80.8   | 15.5  |
| **Proposed**                                                           |                                            | 43.4   | 72.3   | 81.2   | 14.8  |

In addition to the comparison against recent state-of-the-art methods in Table 1 for video retrieval on MSR-VTT, in Table 5 we show the complete comparison to prior work and summarize the models here. Collaborative Experts [26] leverages “expert” features that can be obtained from the raw video from different off-the-shelf models (such as object detection, scene classification, and speech recognition models) to build representations. Instead of summarizing the expert features into a compact video representation and computing similarity with the text representation, the Multi-Modal Transformer [10] computes similarity between different expert features and the text representation with a proposed variation of the Transformer [42]. Based on the Multi-Modal Transformer, Multidomain Multi-Modal Transformer [9] explored an additional motion feature and the combination of different training datasets to further improve the result. Support-Set Bottlenecks [35] studies the benefit that cross-instance captioning can bring by generating text based on the combination of all rep-
resentations of similar videos. Similar to our framework, Hierarchical Transformer with Momentum Contrast [24] divided representations from different layer of the encoders into fine-grained (which they referred to feature-level) and high-level (which they referred to semantic-level) representations. While our work focused on learning discrete representations in the fine-grained embedding space, they performed momentum-based representation matching across the two levels that encourages two embedding spaces to be more similar. TeachText [5] leverages distillation learning where multiple captions describing the same video can be considered by different teacher models that jointly guide the student network. Frozen-in-Time [2] and CLIP4Clip [27] both found the recent proposed Visual Transformer [8] can significantly improve retrieval results while they differ in the choice of summarizing function for video (i.e. $f_{\text{high}}^M$) and the pre-training procedure. Moreover, CLIP4Clip also introduces different choice of the summarizing function $f_{\text{high}}^M$, including RNNs (CLIP4Clip-seqLSTM) and Transformers (CLIP4Clip-seqTransformer) that replaces the mean-pooling function (CLIP4Clip-meanPooling) at the cost of higher time complexity and computational cost. Note that while our work is based on the vanilla mean-pooling function, we achieved comparable or better performance with the proposed discrete embedding representations. As described in Section 3, CLIP4Clip also introduced a cross-modal transformer network (CLIP4Clip-tightTransformer) that allows cross-modal reference for deriving representations.

### D Places Image Retrieval Full Comparison

Table 6: Full comparison against prior works on Places image and spoken caption retrieval.

|                      | Audio to Image | Image to Audio |
|----------------------|----------------|----------------|
|                      | R@1 ↑ | R@5 ↑ | R@10 ↑ | MnR ↓ | R@1 ↑ | R@5 ↑ | R@10 ↑ | MnR ↓ |
| Harwath et al. [17]‡ | 14.8  | 40.3  | 54.8  | –     | 12.1  | 33.5  | 46.3  | –     |
| Harwath et al. [13]‡ | 16.1  | 40.4  | 56.4  | –     | 13.0  | 37.8  | 54.2  | –     |
| DAVEEnet [15]        | 20.0  | 46.9  | 60.4  | –     | 12.7  | 37.5  | 52.8  | –     |
| ResDAVEEnet [16]*    | 30.9  | 63.6  | 74.2  | 20.2  | 26.4  | 58.5  | 71.2  | 21.6  |
| ResDAVEEnet-VQ [14]* | 34.9  | 70.2  | 79.4  | 15.0  | 32.7  | 65.6  | 77.0  | 18.0  |
| Our Baseline†        | 43.8  | 74.1  | 82.4  | 15.8  | 40.4  | 73.3  | 82.5  | 10.9  |
| **Proposed**         | **46.5** | **77.4** | **85.8** | **13.7** | **45.4** | **77.7** | **85.9** | **8.9** |

‡ Results found in [15].
† Existing model reproduced with $\mathcal{L}_{\text{MMS}}$ for fair comparison.
* Results obtained by running the official code and pre-trained models.

We show the full comparison to prior work on Places-400k in Table 6. The previous methods [13, 17, 15] use less complex audio and image encoders with fewer parameters.

### E Results Without Cross-Modal Code Matching

To demonstrate the importance of our proposed Cross-Modal Code Matching objective, Figure 5 illustrates the conditional probability matrix (described in Section 4.3 and Figure 3) when the proposed objective is deactivated (setting $\alpha = 0$). Unsurprisingly, we see that the correlation between codewords and action labels are gone, indicating that the assignment of codewords are now dominated by the input modality instead of the underlying action label. This can also be verified by visualizing the discrete embedding space in a lower dimension as plotted in Figure 6. This evidence suggests that the proposed Cross-Modal Code Matching Objective is effective for learning modality-invariant representations.

### F Additional Codeword Correspondence and Localization Examples

An extension of Table 2 showing the correspondence between codewords, visual actions, and spoken words are provided in Table 7. We also provide more examples for codeword localization in Figure 7.
Figure 5: Conditional probability matrix between codewords and action labels learned by our proposed method when the Cross-Modal Code Matching objective is excluded.

Figure 6: T-SNE visualization of the codebook with and without the proposed Cross-Modal Code Matching Objective. Each point corresponds to a codeword colored with respect to the input modality that utilized it the most. Codewords without high (> 90%) usage from single modality are labeled as “jointly used”.

Figure 7: More examples for codeword cross-modal localization.
Table 7: Correspondence between codewords, visual actions, and spoken words (Extended Table 2). The second hypothesis and the occurrence are omitted for simplicity. All codewords activated on S-MiT’s development set are listed.

| Rank | Code | Visual Action | Spoken word | Rank | Code | Visual Action | Spoken word |
|------|------|---------------|-------------|------|------|---------------|-------------|
| 1    | 201  | juggling      | 97.5        | 61   | 940  | landing       | 44.9        |
| 2    | 349  | flossing      | 96.0        | 62   | 262  | sewing        | 44.7        |
| 3    | 145  | surfing       | 95.6        | 63   | 532  | autographing  | 44.4        |
| 4    | 29   | tattooing     | 94.6        | 64   | 928  | stirring      | 44.1        |
| 5    | 233  | ironing       | 93.8        | 65   | 747  | applauding    | 43.8        |
| 6    | 766  | surfing       | 93.2        | 66   | 447  | paddling      | 43.1        |
| 7    | 191  | juggling      | 90.2        | 67   | 823  | skipping      | 43.0        |
| 8    | 753  | autographing  | 85.0        | 68   | 308  | shaving       | 42.5        |
| 9    | 606  | autographing  | 83.7        | 69   | 518  | skiing        | 41.8        |
| 10   | 640  | drumming      | 81.6        | 70   | 860  | bulldozing    | 41.7        |
| 11   | 436  | injecting     | 81.6        | 71   | 616  | extinguishing | 41.3        |
| 12   | 109  | peeling       | 80.9        | 72   | 296  | combing       | 40.9        |
| 13   | 551  | shaving       | 80.2        | 73   | 435  | paddling      | 38.6        |
| 14   | 137  | paddling      | 80.0        | 74   | 705  | surfing       | 38.0        |
| 15   | 327  | crying        | 78.8        | 75   | 760  | hammering     | 38.0        |
| 16   | 593  | surfing       | 77.7        | 76   | 926  | paddling      | 37.5        |
| 17   | 687  | drumming      | 77.3        | 77   | 888  | paddling      | 37.1        |
| 18   | 103  | tattooing     | 77.2        | 78   | 169  | dunking       | 36.9        |
| 19   | 1000 | inflating     | 74.5        | 79   | 681  | maneuvering   | 36.7        |
| 20   | 222  | boxing        | 71.3        | 80   | 685  | signing       | 36.5        |
| 21   | 243  | shredding     | 70.0        | 81   | 631  | paddling      | 36.4        |
| 22   | 157  | paddling      | 69.9        | 82   | 800  | dropping      | 36.3        |
| 23   | 427  | boxing        | 69.8        | 83   | 556  | drumming      | 35.8        |
| 24   | 774  | surfing       | 69.2        | 84   | 758  | wrapping      | 35.7        |
| 25   | 613  | manicuring    | 67.9        | 85   | 368  | texting       | 35.6        |
| 26   | 952  | leaping       | 66.0        | 86   | 625  | combing       | 35.6        |
| 27   | 196  | boxing        | 64.1        | 87   | 166  | boxing        | 35.5        |
| 28   | 706  | sailing       | 63.4        | 88   | 539  | paddling      | 35.4        |
| 29   | 58   | shaving       | 62.8        | 89   | 139  | leaping       | 35.4        |
| 30   | 759  | paddling      | 60.7        | 90   | 123  | drumming      | 35.3        |
| 31   | 868  | boxing        | 60.0        | 91   | 577  | drumming      | 35.3        |
| 32   | 500  | dialing       | 58.9        | 92   | 780  | scrubbing     | 35.2        |
| 33   | 536  | cheering      | 58.6        | 93   | 621  | leaping       | 35.2        |
| 34   | 50   | rafting       | 58.3        | 94   | 154  | boxing        | 35.1        |
| 35   | 664  | dunking       | 58.0        | 95   | 415  | grilling      | 35.1        |
| 36   | 103  | autographing  | 57.8        | 96   | 345  | autographing  | 35.0        |
| 37   | 990  | wrestling     | 56.1        | 97   | 694  | sleeping      | 34.9        |
| 38   | 880  | sleeping      | 56.0        | 98   | 973  | leaping       | 34.4        |
| 39   | 48   | paddling      | 55.1        | 99   | 957  | shrugging     | 34.3        |
| 40   | 292  | skiing        | 54.2        | 100  | 713  | paddling      | 34.2        |
| 41   | 602  | ironing       | 52.5        | 101  | 699  | injecting     | 34.1        |
| 42   | 954  | dropping      | 52.4        | 102  | 431  | peeling       | 33.9        |
| 43   | 735  | applauding    | 52.1        | 103  | 164  | typing        | 33.8        |
| 44   | 816  | autographing  | 51.0        | 104  | 776  | juggling      | 33.6        |
| 45   | 516  | swinging      | 50.0        | 105  | 73   | shrugging     | 32.9        |
| 46   | 421  | carving       | 50.0        | 106  | 846  | injecting     | 32.8        |
| 47   | 168  | drumming      | 49.3        | 107  | 395  | juggling      | 32.7        |
| 48   | 561  | flossing      | 48.0        | 108  | 273  | dusting       | 32.6        |
| 49   | 970  | marrying      | 47.8        | 109  | 737  | puddle        | 32.5        |
| 50   | 610  | dunking       | 47.4        | 110  | 291  | coughing      | 32.4        |
| 51   | 105  | paddling      | 47.2        | 111  | 375  | colliding     | 32.4        |
| 52   | 150  | waxing        | 47.2        | 112  | 693  | sleeping      | 32.3        |
| 53   | 92   | bowing        | 46.7        | 113  | 111  | baking        | 32.3        |
| 54   | 529  | typing        | 46.3        | 114  | 805  | massaging     | 32.0        |
| 55   | 844  | drumming      | 46.2        | 115  | 134  | autographing  | 31.7        |
| 56   | 997  | cheering      | 45.8        | 116  | 923  | wrapping      | 31.6        |
| 57   | 322  | paddling      | 45.8        | 117  | 698  | surfing       | 31.5        |
| 58   | 672  | boxing        | 45.6        | 118  | 362  | paddling      | 31.5        |
| 59   | 97   | barbecuing    | 45.6        | 119  | 505  | drumming      | 31.0        |
| 60   | 216  | inflating     | 45.3        | 120  | 215  | shaving       | 31.0        |
Table 6: continued

| Rank | Code | Visual Action | Spoken word | Rank | Code | Visual Action | Spoken word |
|------|------|---------------|-------------|------|------|---------------|-------------|
| 121  | 642  | autographing  | 30.8        | 181  | 646  | autographing  | 23.5        |
| 122  | 828  | paddling      | 30.6        | 182  | 423  | applauding    | 23.5        |
| 123  | 6    | leaping       | 30.3        | 183  | 699  | racing        | 23.4        |
| 124  | 974  | sprinkling    | 30.0        | 184  | 651  | paddling      | 23.4        |
| 125  | 44   | flossing      | 29.9        | 185  | 414  | drenching     | 23.3        |
| 126  | 324  | drumming      | 29.9        | 186  | 55   | racing        | 23.3        |
| 127  | 108  | boxing        | 29.8        | 187  | 718  | drumming      | 23.2        |
| 128  | 784  | pedaling      | 29.7        | 188  | 439  | pedaling      | 23.1        |
| 129  | 266  | barbecuing    | 29.7        | 189  | 255  | paddling      | 23.1        |
| 130  | 991  | drumming      | 29.6        | 190  | 77   | clipping      | 23.1        |
| 131  | 597  | signing       | 29.3        | 191  | 701  | lecturing     | 23.0        |
| 132  | 817  | welding       | 29.1        | 192  | 19   | clipping      | 22.9        |
| 133  | 673  | typing        | 29.1        | 193  | 859  | singing       | 22.9        |
| 134  | 113  | dialing       | 29.0        | 194  | 444  | autographing  | 22.9        |
| 135  | 470  | sawing        | 28.9        | 195  | 371  | barbecuing    | 22.9        |
| 136  | 657  | landing       | 28.7        | 196  | 315  | peeling       | 22.8        |
| 137  | 440  | surfing        | 28.6        | 197  | 271  | racing        | 22.7        |
| 138  | 404  | rinsing        | 28.6        | 198  | 955  | leaping       | 22.6        |
| 139  | 0    | applauding    | 28.6        | 199  | 584  | boxing        | 22.6        |
| 140  | 950  | paddling      | 28.2        | 200  | 555  | pitching      | 22.5        |
| 141  | 430  | hiking        | 27.8        | 201  | 286  | piloting      | 22.5        |
| 142  | 762  | leaping       | 27.8        | 202  | 569  | paddling      | 22.3        |
| 143  | 504  | bowling        | 27.3        | 203  | 692  | paddling      | 22.2        |
| 144  | 295  | pedaling      | 27.3        | 204  | 682  | paddling      | 22.1        |
| 145  | 579  | dunking       | 27.2        | 205  | 116  | sequencing    | 22.0        |
| 146  | 380  | leaping       | 26.7        | 206  | 442  | dropping      | 22.0        |
| 147  | 445  | sleeping       | 26.7        | 207  | 324  | skiing        | 22.0        |
| 148  | 603  | leaping       | 26.5        | 208  | 924  | flooding      | 21.9        |
| 149  | 838  | dusting        | 26.5        | 209  | 826  | bulldozing    | 21.6        |
| 150  | 825  | scooping       | 25.9        | 210  | 422  | falling       | 21.4        |
| 151  | 64   | pedaling       | 25.9        | 211  | 931  | bulldozing    | 21.4        |
| 152  | 455  | erupting       | 25.6        | 212  | 259  | wrestling     | 21.3        |
| 153  | 429  | competing      | 25.5        | 213  | 475  | leaping       | 21.2        |
| 154  | 989  | competing      | 25.5        | 214  | 905  | leaping       | 21.2        |
| 155  | 223  | competing      | 25.4        | 215  | 806  | leaping       | 21.2        |
| 156  | 51   | bowling        | 25.4        | 216  | 813  | leaping       | 21.1        |
| 157  | 379  | slicing        | 25.4        | 217  | 538  | paddling      | 21.0        |
| 158  | 911  | paddling       | 25.4        | 218  | 101  | sanding       | 20.9        |
| 159  | 364  | leaping        | 25.4        | 219  | 482  | swinging      | 20.9        |
| 160  | 483  | paddling       | 25.3        | 220  | 680  | leaping       | 20.9        |
| 161  | 634  | autographing   | 25.0        | 221  | 1018 | dialing       | 20.7        |
| 162  | 884  | leaping        | 25.0        | 222  | 665  | shaving       | 20.7        |
| 163  | 485  | stirring       | 25.0        | 223  | 417  | leaping       | 20.6        |
| 164  | 540  | boxing         | 25.0        | 224  | 165  | leaping       | 20.6        |
| 165  | 13   | paddling       | 25.0        | 225  | 194  | leaping       | 20.6        |
| 166  | 873  | paddling       | 25.0        | 226  | 200  | leaping       | 20.5        |
| 167  | 909  | autographing   | 24.3        | 227  | 453  | leaping       | 20.4        |
| 168  | 638  | autographing   | 24.3        | 228  | 57   | singing       | 20.2        |
| 169  | 963  | plugging       | 24.3        | 229  | 420  | paddling      | 20.0        |
| 170  | 131  | paddling       | 24.2        | 230  | 918  | massaging     | 19.8        |
| 171  | 799  | welding        | 24.2        | 231  | 810  | paddling      | 19.8        |
| 172  | 486  | hammering      | 24.1        | 232  | 520  | sailing       | 19.7        |
| 173  | 465  | competing      | 24.0        | 233  | 190  | knitting      | 19.6        |
| 174  | 6    | lecturing      | 24.0        | 234  | 1016 | mopping       | 19.6        |
| 175  | 325  | texting        | 24.0        | 235  | 317  | leaping       | 19.4        |
| 176  | 1001 | competing      | 23.9        | 236  | 827  | paddling      | 19.3        |
| 177  | 242  | competing      | 23.9        | 237  | 24   | leaping       | 19.2        |
| 178  | 714  | calling        | 23.7        | 238  | 1019 | leaping       | 19.1        |
| 179  | 89   | competing      | 23.6        | 239  | 997  | sleeping      | 19.0        |
| 180  | 1013 | paddling       | 23.5        | 240  | 77   | peeling       | 19.0        |
Table 6: continued

| Rank | Code | Visual Action | Spoken word | Visual Action | Spoken word |
|------|------|---------------|--------------|---------------|--------------|
| 241  | 126  | leaping       | 18.9         | exercising    | 18.8         |
| 242  | 449  | leaping       | 18.9         | riding        | 12.3         |
| 243  | 187  | surfing       | 18.9         | walking       | 13.3         |
| 244  | 117  | riding        | 18.8         | traffic       | 21.7         |
| 245  | 671  | paddling      | 18.8         | city          | 13.8         |
| 246  | 736  | autographing  | 18.7         | howling       | 4.9          |
| 247  | 251  | surfing       | 18.5         | tree          | 17.6         |
| 248  | 491  | raining       | 18.4         | simpsons      | 8.5          |
| 249  | 1     | burying       | 18.4         | dirt          | 19.3         |
| 250  | 188  | paddling      | 18.4         | beard         | 8.8          |
| 251  | 742  | pedaling      | 18.3         | bike          | 22.0         |
| 252  | 531  | chewing       | 18.2         | eats          | 12.5         |
| 253  | 130  | applauding    | 18.1         | soldier       | 12.0         |
| 254  | 246  | discussing    | 17.9         | office        | 23.5         |
| 255  | 503  | lecturing     | 17.6         | bringing      | 15.0         |
| 256  | 961  | paddling      | 17.6         | rushing       | 8.3          |
| 257  | 387  | paddling      | 17.6         | dancing       | 15.0         |
| 258  | 531  | paddling      | 17.6         | dancing       | 15.0         |
| 259  | 1008 | paddling      | 17.6         | dancing       | 15.0         |
| 260  | 554  | sewing        | 17.6         | machine       | 12.2         |
| 261  | 604  | paddling      | 17.6         | machine       | 12.2         |
| 262  | 587  | saluting      | 17.5         | soldier       | 12.0         |
| 263  | 509  | discussing    | 17.5         | office        | 23.5         |
| 264  | 720  | competing     | 17.5         | track         | 24.2         |
| 265  | 1022 | shrugging     | 17.5         | gum           | 18.1         |
| 266  | 987  | autographing  | 17.4         | baseball      | 21.3         |
| 267  | 294  | paddling      | 17.4         | stick         | 16.9         |
| 268  | 1008 | paddling      | 17.4         | scuba         | 4.1          |
| 269  | 284  | paddling      | 17.4         | concert       | 29.3         |
| 270  | 288  | paddling      | 17.3         | water         | 5.8          |
| 271  | 587  | paddling      | 17.3         | water         | 5.8          |
| 272  | 720  | paddling      | 17.3         | water         | 5.8          |
| 273  | 720  | paddling      | 17.3         | water         | 5.8          |
| 274  | 1022 | paddling      | 17.3         | water         | 5.8          |
| 275  | 987  | paddling      | 17.3         | water         | 5.8          |
| 276  | 294  | paddling      | 17.3         | water         | 5.8          |
| 277  | 1008 | paddling      | 17.3         | water         | 5.8          |
| 278  | 284  | paddling      | 17.3         | water         | 5.8          |
| 279  | 288  | paddling      | 17.3         | water         | 5.8          |
| 280  | 587  | paddling      | 17.3         | water         | 5.8          |
| 281  | 720  | paddling      | 17.3         | water         | 5.8          |
| 282  | 1022 | paddling      | 17.3         | water         | 5.8          |
| 283  | 987  | paddling      | 17.3         | water         | 5.8          |
| 284  | 294  | paddling      | 17.3         | water         | 5.8          |
| 285  | 1008 | paddling      | 17.3         | water         | 5.8          |
| 286  | 284  | paddling      | 17.3         | water         | 5.8          |
| 287  | 288  | paddling      | 17.3         | water         | 5.8          |
| 288  | 587  | paddling      | 17.3         | water         | 5.8          |
| 289  | 720  | paddling      | 17.3         | water         | 5.8          |
| 290  | 1022 | paddling      | 17.3         | water         | 5.8          |
| 291  | 987  | paddling      | 17.3         | water         | 5.8          |
| 292  | 294  | paddling      | 17.3         | water         | 5.8          |
| 293  | 1008 | paddling      | 17.3         | water         | 5.8          |
| 294  | 284  | paddling      | 17.3         | water         | 5.8          |
| 295  | 288  | paddling      | 17.3         | water         | 5.8          |
| 296  | 587  | paddling      | 17.3         | water         | 5.8          |
| 297  | 720  | paddling      | 17.3         | water         | 5.8          |
| 298  | 1022 | paddling      | 17.3         | water         | 5.8          |
| 299  | 987  | paddling      | 17.3         | water         | 5.8          |
| 300  | 294  | paddling      | 17.3         | water         | 5.8          |
| Rank | Code | Visual Action | Spoken word | Code | Visual Action | Spoken word |
|------|------|---------------|-------------|------|---------------|-------------|
| 361  | 623  | paddling      | 14.5 yellow | 421  | 282          | dropping    |
| 362  | 462  | leaping       | 14.5 dancing | 422  | 443          | frying      |
| 363  | 65   | autographing  | 14.5 pen    | 423  | 676          | rinsing     |
| 364  | 480  | leaping       | 14.5 greetings | 424  | 578          | grilling    |
| 365  | 376  | paddling      | 14.5 large  | 425  | 994          | autographing|
| 366  | 362  | leaping       | 14.5 dancing | 426  | 391          | locking     |
| 367  | 730  | paddling      | 14.4 red    | 427  | 155          | massaging   |
| 368  | 460  | trimming      | 14.3 tomatoes| 428  | 920          | competing   |
| 369  | 861  | dusting       | 14.3 swiffer | 429  | 204          | autographing|
| 370  | 537  | leaping       | 14.3 daughter| 430  | 959          | manicuring  |
| 371  | 933  | towing        | 14.3 truck  | 431  | 896          | tape        |
| 372  | 636  | paddling      | 14.3 trees  | 432  | 320          | shaving     |
| 373  | 336  | leaping       | 14.2 fire   | 433  | 835          | drawing     |
| 374  | 794  | juggling      | 14.2 boy    | 434  | 202          | dropping    |
| 375  | 283  | leaping       | 14.1 statue  | 435  | 983          | rinsing     |
| 376  | 703  | paddling      | 14.4 camera | 436  | 519          | autographing|
| 377  | 419  | leaping       | 14.0 flying | 437  | 945          | lecturing   |
| 378  | 507  | racing        | 14.0 track  | 438  | 754          | paddling    |
| 379  | 445  | driving       | 14.0 cars   | 439  | 601          | dropping    |
| 380  | 11   | crouching     | 14.0 kneeling| 440  | 140          | lecturing   |
| 381  | 74   | autographing  | 13.9 blond  | 441  | 896          | filing      |
| 382  | 901  | singing       | 13.9 girl   | 442  | 408          | paddling    |
| 383  | 313  | leaping       | 13.9 toys   | 443  | 333          | draining    |
| 384  | 346  | packing       | 13.8 conveyor| 444  | 979          | lecturing   |
| 385  | 508  | paddling      | 13.8 person | 445  | 871          | falling     |
| 386  | 267  | saluting      | 13.8 soldiers| 446  | 53           | paddling    |
| 387  | 452  | drumming      | 13.8 stage  | 447  | 32           | paddling    |
| 388  | 944  | massaging     | 13.8 back   | 448  | 488          | pedaling    |
| 389  | 595  | juggling      | 13.8 throws | 449  | 839          | pushing     |
| 390  | 224  | paddling      | 13.8 day    | 450  | 378          | dunking     |
| 391  | 619  | shredding     | 13.8 machinery| 451  | 489          | applauding  |
| 392  | 512  | juggling      | 13.7 t-shirt| 452  | 999          | leaping     |
| 393  | 160  | autographing  | 13.7 paper  | 453  | 339          | skating     |
| 394  | 390  | pouring       | 13.7 liquid | 454  | 653          | dropping    |
| 395  | 394  | paddling      | 13.7 car    | 455  | 225          | autographing|
| 396  | 541  | flossing      | 13.6 fancy  | 456  | 30           | leaping     |
| 397  | 936  | massaging     | 13.6 electronic| 457  | 654          | applauding  |
| 398  | 321  | standing      | 13.4 performing| 458  | 102          | tattooing   |
| 399  | 432  | weeding       | 13.4 garden | 459  | 662          | autographing|
| 400  | 71   | bulldozing    | 13.3 tractor| 460  | 219          | talking     |
| 401  | 596  | drenching     | 13.3 window | 461  | 99           | dropping    |
| 402  | 177  | autographing  | 13.3 broadcast| 462  | 669          | shaving     |
| 403  | 10   | dialing       | 13.3 jack   | 463  | 962          | dropping    |
| 404  | 527  | autographing  | 13.3 street | 464  | 205          | sleeping    |
| 405  | 837  | drenching     | 13.3 rain   | 465  | 936          | dropping    |
| 406  | 293  | leaping       | 13.3 fly    | 466  | 728          | applauding  |
| 407  | 867  | dropping      | 13.3 bunch  | 467  | 804          | leaping     |
| 408  | 426  | weeding       | 13.3 gardening| 468  | 529          | leaping     |
| 409  | 280  | leaping       | 13.3 dog    | 469  | 971          | hitchhiking |
| 410  | 331  | autographing  | 13.1 contract| 470  | 666          | applauding  |
| 411  | 523  | leaping       | 13.1 dancing| 471  | 797          | applying    |
| 412  | 741  | singing       | 13.0 microphone| 472  | 425          | drumming    |
| 413  | 744  | barbecuing    | 13.0 chef   | 473  | 663          | peeling     |
| 414  | 724  | sawing        | 13.0 tree   | 474  | 471          | applauding  |
| 415  | 277  | juggling      | 13.0 motion | 475  | 711          | leaping     |
| 416  | 16   | dialing       | 12.9 device | 476  | 611          | sleeping    |
| 417  | 984  | destroying    | 12.9 tower  | 477  | 715          | paddling    |
| 418  | 917  | dragging      | 12.9 pulling| 478  | 36           | singing     |
| 419  | 729  | leaping       | 12.8 running| 479  | 700          | tattooing   |
| 420  | 365  | autographing  | 12.8 walk   | 480  | 1017         | applauding  |
| Rank | Code | Visual Action | Spoken word | Prc. word | F1  |
|------|------|---------------|-------------|-----------|-----|
| 481  | 887  | autographing  | 11.4        | sidewalk  | 4.7 |
| 482  | 829  | leaping       | 11.4        | cat       | 14.5|
| 483  | 162  | lecturing     | 11.4        | speaking  | 7.3 |
| 484  | 852  | swimming      | 11.3        | pool      | 15.4|
| 485  | 353  | paddling      | 11.2        | trickling | 2.1 |
| 486  | 998  | paddling      | 11.2        | green     | 22.4|
| 487  | 565  | manicuring    | 11.2        | painting  | 8.6 |
| 488  | 129  | drumming      | 11.2        | night     | 25.9|
| 489  | 21   | paddling      | 11.2        | going     | 7.6 |
| 490  | 690  | autographing  | 11.2        | blond     | 7.2 |
| 491  | 214  | slipping      | 11.1        | snow      | 11.0|
| 492  | 454  | paddling      | 11.1        | bridge    | 5.6 |
| 493  | 320  | unpacking     | 11.1        | boxes     | 18.2|
| 494  | 261  | paddling      | 11.1        | down      | 3.4 |
| 495  | 864  | sleeping      | 11.1        | father    | 7.1 |
| 496  | 411  | burying       | 11.0        | hole      | 16.0|
| 497  | 127  | competing     | 11.0        | field     | 5.2 |
| 498  | 580  | child+singing | 10.8        | girl      | 10.7|
| 499  | 849  | paddling      | 10.8        | slowly    | 8.4 |
| 500  | 285  | autographing  | 10.7        | dress     | 9.6 |
| 501  | 721  | autographing  | 10.7        | middle-aged| 2.6|
| 502  | 88   | leaping       | 10.7        | wall      | 10.1|
| 503  | 769  | autographing  | 10.6        | table     | 10.2|
| 504  | 91   | autographing  | 10.6        | she       | 8.1 |
| 505  | 119  | jumping       | 10.6        | rope      | 8.8 |
| 506  | 448  | paddling      | 10.6        | hat       | 8.3 |
| 507  | 831  | skating       | 10.5        | park      | 20.6|
| 508  | 906  | leaping       | 10.5        | store     | 9.6 |
| 509  | 344  | discussing    | 10.5        | restaurant| 25.7|
| 510  | 847  | cheering      | 10.5        | competition| 4.3|
| 511  | 357  | shaving       | 10.4        | his       | 5.3 |
| 512  | 904  | running       | 10.4        | running   | 13.1|
| 513  | 193  | paddling      | 10.4        | someone   | 16.2|
| 514  | 192  | applauding    | 10.3        | motocross | 6.2 |
| 515  | 230  | autographing  | 10.3        | looking   | 10.8|
| 516  | 534  | sleeping      | 10.3        | bag       | 4.5 |
| 517  | 550  | peeling       | 10.2        | bowl      | 15.7|
| 518  | 159  | autographing  | 10.2        | ward      | 4.3 |
| 519  | 314  | leaping       | 10.2        | mixed-race| 4.0 |
| 520  | 709  | leaping       | 10.1        | animals   | 4.8 |
| 521  | 95   | sprinkling     | 10.1        | sprinkler | 10.8|
| 522  | 227  | sleeping      | 10.1        | oh        | 2.7 |
| 523  | 935  | applauding    | 10.1        | perch     | 12.5|
| 524  | 176  | typing        | 10.1        | office    | 5.3 |
| 525  | 1011 | drumming      | 10.0        | boy       | 17.0|
| 526  | 683  | competing     | 10.0        | game      | 7.1 |
| 527  | 185  | knitting      | 10.0        | stitching | 8.2 |
| 528  | 289  | dropping      | 10.0        | ground    | 16.5|
| 529  | 899  | reaching      | 10.0        | church    | 20.8|
| 530  | 767  | playing       | 10.0        | overwatch | 6.7 |
| 531  | 796  | paddling      | 10.0        | base      | 3.7 |
| 532  | 161  | discussing    | 9.9         | family    | 10.8|
| 533  | 782  | leaping       | 9.9         | doing     | 12.9|
| 534  | 850  | autographing  | 9.9         | american | 1.9 |
| 535  | 620  | leaping       | 9.8         | bridge    | 4.1 |
| 536  | 992  | leaping       | 9.8         | point-of-view | 6.9|
| 537  | 547  | grilling      | 9.8         | crawling | 17.3|
| 538  | 891  | paddling      | 9.8         | on       | 2.5 |
| 539  | 340  | dusting       | 9.8         | clean.    | 5.2 |
| 540  | 659  | storming      | 9.8         | yard     | 18.2|

Table 6: continued
| Rank | Code | Visual Action | Spoken word | Rank | Code | Visual Action | Spoken word |
|------|------|---------------|-------------|------|------|---------------|-------------|
| 601  | 723  | autographing  | 8.7 something | 661  | 980  | kicking       | 7.7 shooting |
| 602  | 719  | talking       | 8.7 toddler   | 662  | 238  | sleeping      | 7.7 squirrel |
| 603  | 209  | autographing  | 8.7 hair      | 663  | 360  | injecting     | 7.7 person   |
| 604  | 521  | rafting       | 8.7 people    | 664  | 853  | camping       | 7.7 tent     |
| 605  | 98   | applauding    | 8.6 stand     | 665  | 652  | autographing  | 7.6 single   |
| 606  | 969  | leaping       | 8.6 kids      | 666  | 893  | watering      | 7.6 watering |
| 607  | 770  | flossing      | 8.6 explaining| 667  | 546  | piloting      | 7.6 lyrics   |
| 608  | 616  | leaping       | 8.5 snow      | 668  | 674  | applauding    | 7.6 night    |
| 609  | 410  | erupting      | 8.5 explodes  | 669  | 881  | autographing  | 7.5 table    |
| 610  | 12   | paddling      | 8.5 distance  | 670  | 236  | hammering     | 7.5 wooden   |
| 611  | 750  | flossing      | 8.5 drinking  | 671  | 778  | leaping       | 7.5 house    |
| 612  | 1009 | autographing  | 8.5 street    | 672  | 260  | flossing      | 7.5 smiling  |
| 613  | 463  | slicing       | 8.5 pieces    | 673  | 399  | paddling      | 7.5 of       |
| 614  | 843  | autographing  | 8.5 speaking  | 674  | 146  | autographing  | 7.5 fans     |
| 615  | 772  | paddling      | 8.5 workers   | 675  | 745  | autographing  | 7.5 language |
| 616  | 781  | leaping       | 8.5 involving | 676  | 207  | paddling      | 7.5 man      |
| 617  | 757  | flossing      | 8.4 caption   | 677  | 732  | smell         | 7.5 flowers  |
| 618  | 791  | pointing      | 8.3 gameplay  | 678  | 647  | autographing  | 7.4 smiles   |
| 619  | 403  | racing        | 8.3 car       | 679  | 893  | autographing  | 7.4 plastic  |
| 620  | 988  | clipping      | 8.3 shoe      | 680  | 416  | drumming      | 7.4 gloves  |
| 621  | 502  | paddling      | 8.3 going     | 681  | 492  | autographing  | 7.4 fans     |
| 622  | 343  | paddling      | 8.3 over      | 682  | 467  | drumming      | 7.4 child    |
| 623  | 450  | shaving       | 8.3 chef      | 683  | 573  | wrapping      | 7.4 box      |
| 624  | 765  | paddling      | 8.2 gymnast   | 684  | 381  | autographing  | 7.4 he       |
| 625  | 476  | paddling      | 8.2 trees     | 685  | 855  | autographing  | 7.3 gentleman|
| 626  | 818  | autographing  | 8.2 vest      | 686  | 939  | peeling       | 7.3 close    |
| 627  | 746  | autographing  | 8.2 street    | 687  | 494  | peeling       | 7.3 hands    |
| 628  | 122  | paddling      | 8.2 people    | 688  | 575  | paddling      | 7.3 a        |
| 629  | 275  | leaping       | 8.2 workout   | 689  | 581  | smashing      | 7.3 building |
| 630  | 592  | hammering     | 8.2 piles     | 690  | 142  | stopping      | 7.3 characters|
| 631  | 1003 | paddling      | 8.2 around    | 691  | 599  | autographing  | 7.3 two      |
| 632  | 7    | paddling      | 8.1 and       | 692  | 309  | paddling      | 7.2 shooting |
| 633  | 257  | raining       | 8.1 blown     | 693  | 264  | drumming      | 7.2 bedroom  |
| 634  | 170  | leaping       | 8.1 running   | 694  | 919  | autographing  | 7.2 hands    |
| 635  | 341  | flossing      | 8.1 how       | 695  | 47   | autographing  | 7.2 woman    |
| 636  | 354  | sewing        | 8.1 machine   | 696  | 570  | paddling      | 7.1 we       |
| 637  | 600  | paddling      | 8.1 each      | 697  | 965  | paddling      | 7.1 red      |
| 638  | 677  | rolling       | 8.1 cooks     | 698  | 361  | autographing  | 7.1 upright  |
| 639  | 133  | sleeping      | 8.1 string    | 699  | 934  | peeling       | 7.1 putting  |
| 640  | 678  | leaping       | 8.0 tree      | 700  | 594  | sitting       | 7.1 inject   |
| 641  | 466  | injecting     | 8.0 close     | 701  | 186  | draining      | 7.1 house    |
| 642  | 1006 | stirring      | 8.0 pot       | 702  | 809  | paddling      | 7.1 man      |
| 643  | 212  | crying        | 8.0 help      | 703  | 783  | paddling      | 7.1 with     |
| 644  | 889  | autographing  | 8.0 young     | 704  | 151  | autographing  | 7.0 store    |
| 645  | 437  | manicuring    | 7.9 fingers   | 705  | 474  | paddling      | 7.0 person   |
| 646  | 468  | jumping       | 7.9 motorcycle| 706  | 1021 | paddling      | 7.0 decorated|
| 647  | 633  | applauding    | 7.9 show      | 707  | 359  | paddling      | 7.0 picture  |
| 648  | 59   | drumming      | 7.9 watching  | 708  | 740  | applauding    | 6.9 people   |
| 649  | 814  | peeling       | 7.9 someone   | 709  | 614  | autographing  | 6.9 knock-knock|
| 650  | 441  | leaping       | 7.9 jeans     | 710  | 237  | leaping       | 6.9 inflating|
| 651  | 775  | inflating     | 7.9 chair     | 711  | 567  | autographing  | 6.9 and      |
| 652  | 46   | autographing  | 7.9 woman     | 712  | 135  | signing       | 6.9 sitting  |
| 653  | 824  | autographing  | 7.9 field     | 713  | 506  | drumming      | 6.9 guys     |
| 654  | 968  | dusting       | 7.9 floor     | 714  | 28   | autographing  | 6.9 hey      |
| 655  | 544  | reaching      | 7.8 climbing  | 715  | 456  | paddling      | 6.9 animated |
| 656  | 389  | autographing  | 7.8 front     | 716  | 768  | autographing  | 6.9 holding  |
| 657  | 819  | lecturing     | 7.8 laughing  | 717  | 879  | leaping       | 6.9 exercising|
| 658  | 498  | shaving       | 7.8 pink      | 718  | 628  | drumming      | 6.9 men      |
| 659  | 9    | paddling      | 7.7 green     | 719  | 80   | discussing    | 6.9 women    |
| 660  | 927  | singing       | 7.7 saying    | 720  | 586  | applauding    | 6.8 haired   |
| Rank | Code | Visual Action | Top Hypothesis label | Prc. word | Spoken word | F1 | Rank | Code | Visual Action | Top Hypothesis label | Prc. word | Spoken word | F1 |
|------|------|---------------|----------------------|----------|-------------|----|------|------|---------------|----------------------|----------|-------------|----|
| 721  | 1004 | pouring       | 6.8                  | color    | 6.8         |    | 734  | 167 | peeling     | 6.7                  | into     | 3.5         |    |
| 722  | 72   | injecting     | 6.8                  | mixing   | 4.7         |    | 735  | 428 | leaping     | 6.7                  | doing    | 2.2         |    |
| 723  | 876  | autographing  | 6.8                  | arena    | 3.8         |    | 736  | 300 | paddling    | 6.6                  | is       | 2.3         |    |
| 724  | 878  | gambling      | 6.8                  | game     | 9.4         |    | 737  | 786 | paddling    | 6.6                  | background | 1.7         |    |
| 725  | 114  | paddling      | 6.8                  | man      | 15.7        |    | 738  | 272 | leaping     | 6.6                  | truck    | 8.5         |    |
| 726  | 764  | raising       | 6.8                  | trash    | 10.0        |    | 739  | 297 | applauding  | 6.6                  | rustic   | 4.5         |    |
| 727  | 326  | leaping       | 6.8                  | zooming  | 3.1         |    | 740  | 303 | paddling    | 6.6                  | down     | 1.0         |    |
| 728  | 217  | surfing       | 6.8                  | surfer   | 2.0         |    | 741  | 93  | paddling    | 6.5                  | car      | 3.3         |    |
| 729  | 107  | pouring       | 6.8                  | bottle   | 5.8         |    | 742  | 542 | baking      | 6.5                  | cupcake  | 9.2         |    |
| 730  | 749  | injecting     | 6.7                  | person’s | 2.2         |    | 743  | 733 | discussing  | 6.5                  | men      | 7.3         |    |
| 731  | 612  | autographing  | 6.7                  | disgust  | 4.8         |    | 744  | 23  | leaping     | 6.5                  | cross    | 5.5         |    |
| 732  | 834  | autographing  | 6.7                  | outside | 12.7        |    | 745  | 104 | injecting   | 6.5                  | beard    | 2.6         |    |
| 733  | 655  | marrying      | 6.7                  | couple   | 7.5         |    | 746  | 279 | leaping     | 6.5                  | garden   | 4.4         |    |