Human Action CLIPs: Detecting AI-Generated Human Motion

Matyas Bohacek[1,2] , **Hany Farid**[3]
[1]Google [2]Stanford University [3]University of California, Berkeley
maty@stanford.edu, hfarid@berkeley.edu

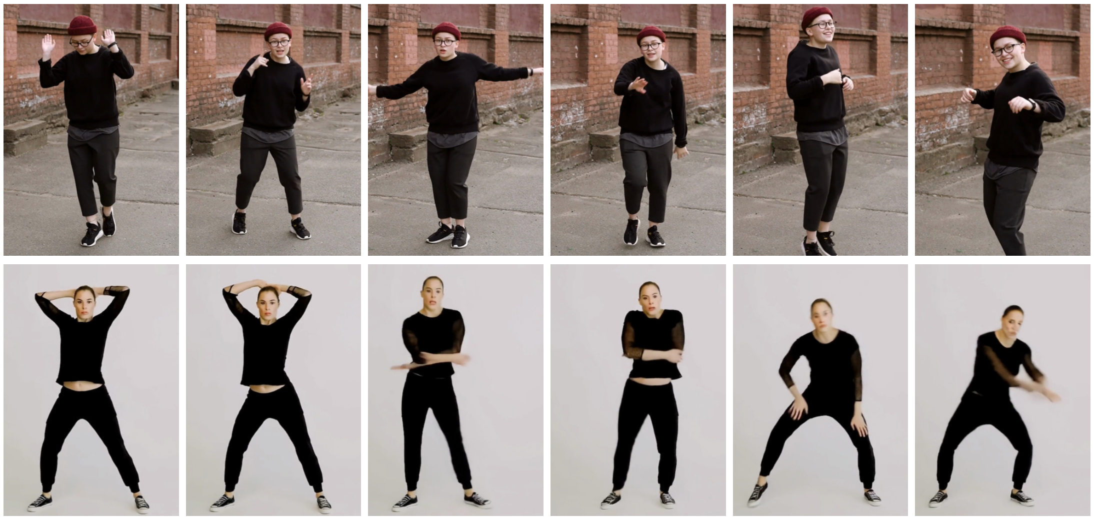

Figure 1: Real video frames sourced from Pexels (Top); AI-generated frames by Veo [Veo, 2024] prompted with "a person doing a dance move" (Bottom). We describe how a low-dimensional CLIP embedding effectively and robustly distinguishes between real and AI-generated videos.

Abstract

AI-generated video generation continues its journey through the uncanny valley to produce content that is increasingly perceptually indistinguishable from reality. To better protect individuals, organizations, and societies from its malicious applications, we describe an effective and robust technique for distinguishing real from AI-generated human motion using multi-modal semantic embeddings. Our method is robust to the types of laundering that typically confound more low- to mid-level approaches, including resolution and compression attacks. This method is evaluated against *DeepAction*, a custom-built, open-sourced dataset of video clips with human actions generated by seven text-to-video AI models and matching real footage. The dataset is available under an academic license at https://www.huggingface.co/datasets/faridlab/deepaction_v1.

1 Introduction

Generating human motion in computer-graphics animation is notoriously difficult because of the complexity of human dynamics and kinematics [McDonnell *et al.*, 2012; Debarba *et al.*, 2020; Diel *et al.*, 2021; Ng *et al.*, 2024] and because of the sensitivity of the human visual system to biological motion [Neri *et al.*, 1998]. While motion capture has significantly improved the realism of complex human motion, gaps have remained.

Generative AI, unlike earlier model-based animation, has emerged as an intriguing new genre for computer animation. While earlier versions of AI-generated video were things nightmares are made of (see, for example, "Will Smith eating spaghetti"[1]), recent advances have shown significant improvements in photo-realism and temporal consistency.

A recent study found that AI-generated faces are nearly indistinguishable from real faces [Nightingale and Farid, 2022], and AI-generated voices are a close second in terms of natu-

[1] https://www.youtube.com/watch?v=XQr4Xklqzw8

ralness and identity [Barrington and Farid, 2024]. There is reason to believe, therefore, that AI-generated videos may pass through the uncanny valley.

We have started to see AI-generated images weaponized in the form of child sexual abuse material, non-consensual sexual imagery, fraud, and as an accelerant to disinformation campaigns [Farid, 2022]. There is no reason, therefore, to believe that AI-generated video will not follow suit.

Video deepfakes fall into two broad categories: impersonation and text-to-video. Although there are several different incarnations of impersonation deepfakes, two of the most popular are lip-sync and face-swap deepfakes. In a face-swap deepfake, a person's face in an original video is replaced with another [Nirkin *et al.*, 2019], and in a lip-sync deepfake, a person's mouth region is modified to be consistent with a new voice track [Suwajanakorn *et al.*, 2017].

By contrast, text-to-video deepfakes are generated entirely from scratch to match a user-specified text prompt. They represent a natural evolution of text-to-image models (e.g., DALL-E, Firefly, Midjourney, etc.). Our focus is on detecting these text-to-video deepfakes, particularly those depicting human motion.

Motivated by, and building upon, earlier work, we describe a technique to detect AI-generated videos containing human motion. Our initial focus is on humans because these videos are of most concern when it comes to the harms enumerated above. Despite this focus, we will discuss why our approach is likely to generalize to other types of video content. We evaluate the efficacy of our approach on a diverse dataset of our creation consisting of real and matching (in terms of the human actions depicted) AI-generated videos from seven different generative-AI models. We demonstrate the robustness of our detection in the face of standard laundering attacks like resizing and transcoding, and evaluate its generalizability to previously unseen models. Our work contributes to this nascent literature with the:

1. development of a task-specific CLIP embedding that outperforms generic CLIP embeddings (FT-CLIP);
2. development of a new unsupervised forensic technique that requires no explicit training (frame-to-prompt);
3. improvement in the generalizability of detection to previously unseen content and synthesis models;
4. extension from image- to video-based analysis; and
5. construction and release of *DeepAction*, a new benchmark dataset of real and AI-generated human motion.

2 Related Work

Identifying manipulated content (image, audio, video) can be partitioned into two broad categories: (1) active and (2) reactive. Active approaches involve inserting metadata or imperceptible watermarks at the time of synthesis to facilitate downstream detection [Collomosse and Parsons, 2024]. These approaches are appealing for their simplicity but are vulnerable to counter-attack in which the inserted credentials can be removed [Voloshynovskiy *et al.*, 2001] (although the extraction and centralized storage of a distinct digital signature – perceptual hash– can be used to reattach credentials).

Reactive techniques – operating in the absence of credentials – fall into two basic approaches: (2a) learning-based, in which features that distinguish real from fake content are learned by a range of machine-learning techniques, and (2b) artifact-based, in which a range of low-level (pixel-based) to high-level (semantic-based) features are explicitly extracted to distinguish between real and fake content [Farid, 2022].

There is a rich literature of techniques for detecting AI-generated images [Farid, 2022] and a more nascent literature for detecting AI-generated voices [Blue *et al.*, 2022; Pianese *et al.*, 2022; Barrington *et al.*, 2023]. The literature for detecting AI-generated or AI-manipulated videos has primarily focused on face-swap deepfakes [Agarwal *et al.*, 2020; Nirkin *et al.*, 2021; Jia *et al.*, 2022], and lip-sync deepfakes [Boháček and Farid, 2022; Bohacek and Farid, 2024; Datta *et al.*, 2024].

Because text-to-video AI generation has only recently emerged as perceptually compelling, the literature on detecting these videos is more sparse. A recent example [Vahdati *et al.*, 2024] leverages low-level features, but these tend to be vulnerable to compression artifacts, which in video – unlike standard image JPEG compression – are highly spatially and temporally variable.

Another recent example [Jia *et al.*, 2024] explores the potential of multi-modal large language models (LLMs) for detecting AI-generated faces.The authors prompt ChatGPT with an image and prompt like "Tell me if this is an AI-generated image." This approach achieves an average accuracy of 75% (as measured by area under the curve, AUC). Although not particularly accurate, what is intriguing about this approach is that it points to a potentially semantic-level reasoning.

Recent studies took a more direct semantic approach by leveraging a contrastive language-image pretraining (CLIP) representation [Radford *et al.*, 2021]. In [Cozzolino *et al.*, 2024], the authors extract a CLIP embedding from real and AI-generated images and with only a linear SVM achieve detection accuracy ranging from 85% to 90% depending on the amount of image post-processing.

In [Khan and Dang-Nguyen, 2024], the authors also exploit CLIP embeddings along with a range of transfer learning strategies to achieve detection accuracy between 95% and 98%; their classifiers show good but not perfect generalizability, achieving an accuracy between 86% and 89%.

3 Dataset

We generated 3,100 video clips from seven text-to-video AI models: BD AnimateDiff [Lin and Yang, 2024], CogVideoX-5B [Yang *et al.*, 2024], Lumiere[2] [Bar-Tal *et al.*, 2024], RunwayML Gen3 [Gen3, 2024], Stable Diffusion Txt2Img+Img2Vid [Blattmann *et al.*, 2023], Veo[3] [Veo, 2024], and VideoPoet [Kondratyuk *et al.*, 2023]. The default generation parameters were used for each model. These AI-generated videos are 297 minutes in length constituting 254,632 video frames, Figure 2.

These video clips depict 100 distinct human actions and vary in length from 2 to 10.7 seconds, in resolution from

[2] An earlier version of the Lumiere model was used which was not specifically trained or fine-tuned on human motion.

[3] A pre-release version of the Veo model was used.

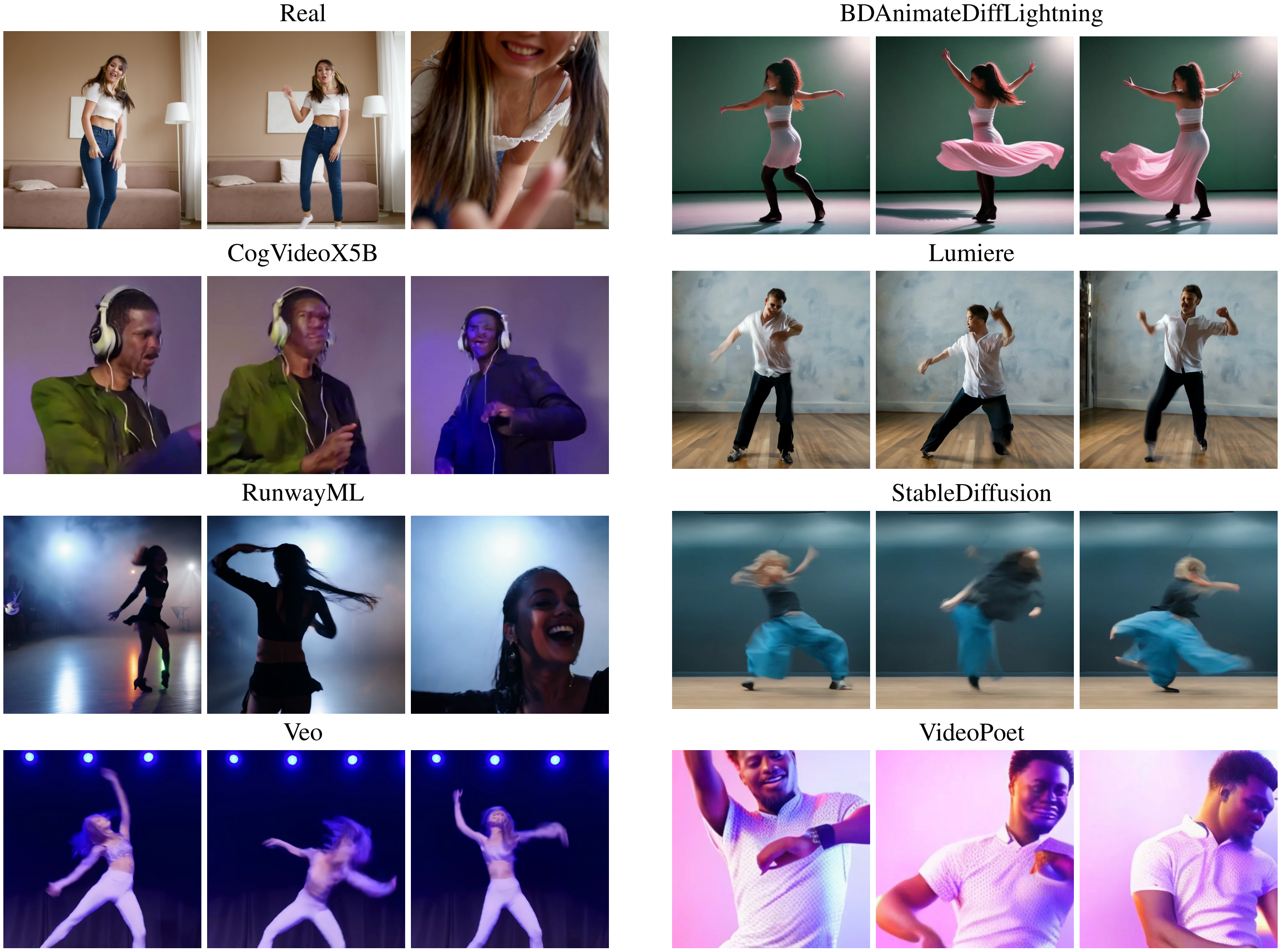

Figure 2: Sample frames from one real and seven AI-generated videos prompted with “a person dancing to music”.

512×512 to 2048×1152 pixels, and in orientation (landscape and portrait). Each video was generated from a short prompt, which itself was generated by asking ChatGPT to create short descriptive prompts of human actions (see Appendix A).

We also curated a set of 100 real videos from Pexels [Pexels, 2024], an open-source stock video database, matched to the human-action prompts described above. These real videos are 28 minutes in length constituting 44,475 individual video frames. These videos were matched on general action so that the embedding representations (described next) between real and fake would not be based on semantic differences.

The resulting dataset, called *DeepAction*, comprises eight video categories (seven AI-generated and one real) and is made publicly available.

4 Methods

We describe four CLIP-like, multi-modal embeddings and proceed with classification schemes used to distinguish the real from the fake.

4.1 Embeddings

Multi-modal embedding models map an image and its corresponding descriptive caption into a shared vector space, allowing comparisons between these modalities. These embeddings have been found to be effective across many computer vision and natural language processing tasks [Shen *et al.*, 2021; Song *et al.*, 2022].

Each video in our dataset is represented as a sequence of video-frame embeddings, extracted using three off-the-shelf embedding models–CLIP [Radford *et al.*, 2021], SigLIP [Zhai *et al.*, 2023], and JinaCLIP [Koukounas *et al.*, 2024]–as well as a custom fine-tuned CLIP model.

1. The **CLIP** embedding model was trained on LAION-2B [Schuhmann *et al.*, 2022]. We used the smallest 512-D baseline model ViT-B/32.

2. The **SigLIP** embedding model was trained on WebLI [Chen *et al.*, 2022], which is over 32% larger than LAION-2B. Unlike CLIP’s softmax-based contrastive loss, SigLIP uses a sigmoid loss. We used the 768-D patch16-224 model variant.

3. The **JinaCLIP** embedding model was trained on LAION-400M [Schuhmann *et al.*, 2021] along with an additional set of 40 text-pair datasets. While CLIP's training optimized for text-image representation alignment, JinaCLIP was trained to jointly optimize text-image and text-text representation alignment. We used the base 768-D v1 model.
4. We created a custom **fine-tuned CLIP** (FT-CLIP) embedding model from the baseline CLIP model described above. We used the same fine-tuning methodology and hyperparameters as described in [Khan and Dang-Nguyen, 2024].

See Appendix for baseline model and training details.

4.2 Classification

We deploy three different classification strategies that, leveraging the embeddings enumerated in the previous section, make a prediction of a video frame being real or fake. The first two supervised classifiers are based on a support vector machine (SVM), and the third unsupervised classifier is based on a simple cosine similarity between text and frame embedding. In each case, a video, represented as a sequence of frame embeddings, is classified as fake if a majority of the frames are classified as fake.

We intentionally take a simple approach here instead of leveraging heavier-weight classifiers so as to place emphasis on the power of the multi-modal embeddings.

1. A **two-class SVM** [Cortes and Vapnik, 1995], implemented in scikit-learn[4], is used to classify video frames as real or fake, in which all seven text-to-video models are bundled into a single 'fake' class.
2. A **multi-class SVM** [Cortes and Vapnik, 1995], also implemented in scikit-learn, classifies the source of each video frame across eight classes. One class represents real videos, and seven classes represent each of seven different text-to-video AI models.
3. The previous classifiers follow a typical supervised learning approach in which the SVMs are trained on a subset of the video frames and evaluated on the remaining frames. In this third **frame-to-prompt** approach, each video-frame embedding is compared – through a simple cosine similarity – to an embedding of one of two prompts (e.g., *"a real image"* and *"a fake image"*). A frame is classified by selecting the class (real/fake) with the largest cosine similarity.

5 Results

We now describe a pair-wise embedding-classifier performance for discriminating between real and AI-generated videos, followed by an evaluation of the robustness in the face of standard laundering attacks, and generalizability to synthesis models not seen during classifier training.

For the CLIP, SigLIP, and JinaCLIP embeddings, our dataset is randomly split into an 80/20 train/test partition.

Embedding	Kernel	Two-Class Frame (%)	Two-Class Video (%)	Multi-Class Frame (%)	Multi-Class Video (%)
CLIP	**linear**	84.3	97.0	87.4	98.2
	RBF	81.6	96.7	89.3	98.3
	poly	79.0	95.9	89.0	98.3
SigLIP	**linear**	89.7	98.1	**90.8**	**98.5**
	RBF	84.4	97.1	**91.5**	**98.5**
	poly	83.0	96.9	**90.9**	**98.4**
JinaCLIP	**linear**	86.9	97.2	85.9	97.2
	RBF	78.1	96.1	87.3	97.9
	poly	76.1	95.3	87.0	97.8
FT-CLIP	**linear**	**90.7**	**98.5**	82.8	97.6
	RBF	**93.2**	**99.1**	84.8	97.6
	poly	**92.6**	**99.2**	85.1	97.7

Table 1: Two-class and multi-class SVM classification accuracy (percent) for four different embeddings (CLIP, SigLIP, JinaCLIP, fine-tuned CLIP) and three different SVM kernels (linear, RBF, polynomial). Results are reported on a per video-frame and per video basis. Across all kernel functions, the fine-tuned CLIP provides the best performance for the two-class model, while SigLIP provides the best performance for the multi-class model.

For the fine-tuned CLIP, the dataset is randomly split into a $40/40/20$ partition, where the first 40% is used for CLIP fine-tuning, 40% is used for model training, and 20% is used for testing. The splits are determined at the action (prompt) level to ensure no overlap across partitions.

In each case, we report the mean frame- and video-level accuracy on the test set, averaged over five random train/test repetitions. Because our dataset is imbalanced, with significantly more fake than real videos, we under-sample the fake videos. Throughout, we report accuracy as a macro-average by evenly weighting the class accuracies.

Two-class. Shown in the left portion of Table 1 is the frame- and video-level accuracy for the two-class SVMs with three different kernels: linear, radial basis function (RBF), and polynomial.

At the frame level, macro-accuracy ranges from a low of 79.0% to a high of 92.6%. For the CLIP embeddings (top three rows), the linear kernel is surprisingly more effective than the non-linear kernels. For the fine-tuned CLIP, the RBF kernel is the highest performer. Although performance is somewhat comparable across different embeddings, the fine-tuned CLIP offers the best performance. At the video level, accuracy across embeddings and classifiers is comparable, ranging from a low of 95.3% to a high of 99.2%.

For all embeddings and all classifiers, there is a slight fake bias in which fake content is correctly classified at a higher rate by, on average, 5.2 percentage points. For example, for the best performing model (FT-CLIP with a poly kernel), the video-level accuracy for fake videos is 99.9% as compared to 98.5% for the real videos.

Shown in Figure 3 is a visualization of seven pairwise, real-fake finetuned-CLIP embeddings, linearly reduced to a 2D subspace using PCA. Even in this reduced space, we see a good separation between the classes (the seven AI models are considered separately only for ease of visualization).

Multi-class. Shown in the right portion of Table 1 is the frame- and video-level accuracy for the multi-class SVMs. At the frame level, accuracy slightly outperforms the two-

[4] https://www.scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

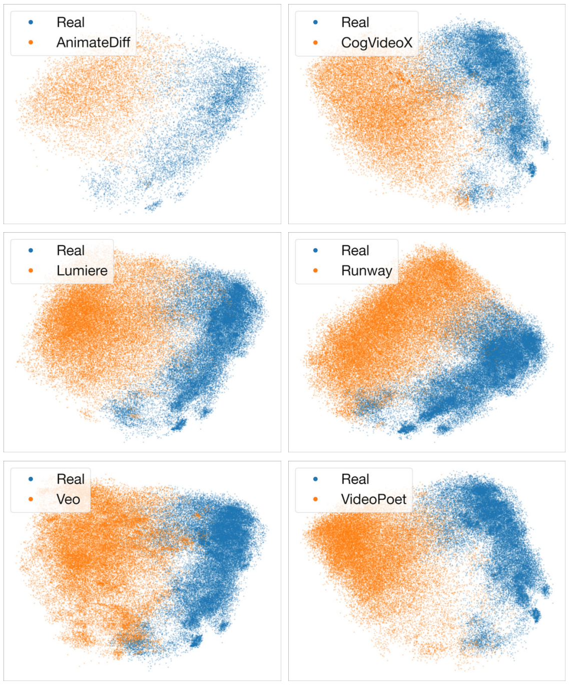

Figure 3: Real (blue) vs fake (orange) fine-tuned CLIP embeddings. See Table 1.

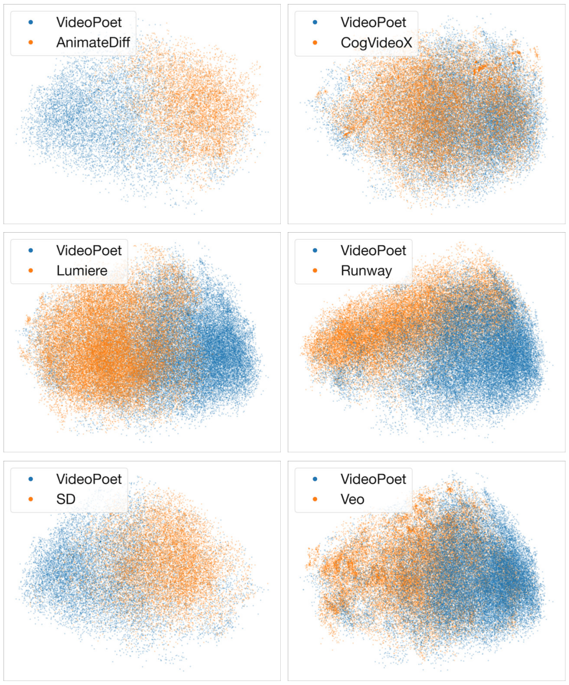

Figure 4: VideoPoet (blue) vs other models (orange) using fine-tuned CLIP embeddings.

class classifier, ranging from a low of 82.8% to a high of 91.5%. Generally speaking, the non-linear kernels (RBF and polynomial) outperform the linear kernel, and unlike the two-class, the SigLIP embedding now outperforms the other embeddings across all kernel functions.

There is a slight bias across all seven AI models with the Veo model consistently misclassified. For example, for the best-performing model (SigLIP with an RBF kernel), the difference between the best (VideoPoet) and worst (Veo) performing inter-class accuracy is 100% and 93.7%.

At the video level, accuracy across all embeddings and classifiers are comparable ranging from a low of 97.6% to a high of 98.5%.

Shown in Figure 4 is a PCA-based visualization of the pairwise fine-tuned CLIP embeddings between one text-to-video model (VideoPoet) and each of six other text-to-video models. Even in this reduced space, we see a good separation between the different AI models (the six AI models are considered separately only for ease of visualization; other pair-wise models exhibit similar patterns).

Frame-to-prompt. Shown in Table 2 is the frame- and video-level accuracy of frame-to-prompt classifiers (Section 4.2) for five different paired prompts:

- (P1) *real photo vs. fake photo*,
- (P2) *real image vs. fake image*,
- (P3) *authentic image vs. AI-generated image*,
- (P4) *authentic photo vs. manipulated photo*,
- (P5) *authentic image vs. manipulated image*.

At the frame level, accuracy is generally significantly lower than the previous two- and multi-class SVMs, with a maximum accuracy of 95.2% for the fine-tuned CLIP embedding and the P1 paired prompt *authentic image vs. AI-generated image*. At the video level, accuracy for the same fine-tuned CLIP and P1 prompt improves slightly to 96.2%. This is perhaps not surprising since we expect these semantic embeddings to be highly correlated across a video.

As with the two-class SVM, we see a bias—only this time, towards real videos. For example, the video-level classification using FT-CLIP and prompt P1 correctly classifies 100% of the real as compared to 92.4% of the fake videos.

Compared to the best-performing two-class SVM with a video-level accuracy of 99.2% (Table 1), the best-performing frame-to-prompt underperforms by only 6.6 percentage points. This is particularly impressive given that the frame-to-prompt approach requires no explicit training.

5.1 Robustness

Whether intentional or not, videos subjected to a forensic analysis often undergo laundering in the form of changes in resolution and compression. Techniques leveraging low-level features are often highly vulnerable to this type of laundering because even these simple modifications obliterate distinguishing characteristics. Because the multi-modal embeddings are designed to extract semantic-level meaning, we expect these features to be more resilient to laundering.

Shown in the left portion of Table 3 are the frame- and video-level accuracies for the CLIP embedding and a two-class linear SVM for videos with progressively lower resolution (we quantify the change in resolution as a percentage

Embedding	Frame (%)				
	P1	P2	P3	P4	P5
CLIP	52.9	48.0	63.6	64.7	60.3
SigLIP	49.0	47.8	49.1	49.1	47.2
JinaCLIP	51.8	55.3	54.0	55.7	58.0
FT-CLIP	**95.2**	91.0	83.1	78.3	79.9
	Video (%)				
	P1	P2	P3	P4	P5
CLIP	53.5	44.2	62.0	63.7	58.4
SigLIP	47.7	46.6	49.2	49.0	47.6
JinaCLIP	56.7	55.6	56.2	59.9	60.5
FT-CLIP	**96.2**	90.7	85.6	80.9	81.8

Table 2: Frame-to-prompt classification accuracy across different embeddings (CLIP, SigLIP, JinaCLIP, fine-tuned CLIP) and five different prompt pairs: (P1) *real photo vs. fake photo*; (P2) *real image vs. fake image*; (P3) *authentic image vs. AI-generated image*; (P4) *authentic photo vs. manipulated photo*; and (P5) *authentic image vs. manipulated image*. Results are reported on a per video-frame and per video basis.

	Resolution		Compression	
amount	Frame (%)	Video (%)	Frame (%)	Video (%)
100%	83.1	96.9	83.1	96.9
75%	78.8	95.0	82.8	96.5
50%	72.5	93.8	81.8	96.5
25%	62.5	72.8	79.5	95.1
10%	49.5	51.0	73.7	92.3

Table 3: Accuracy for a two-class linear SVM and CLIP embeddings for videos at 100%, 75%, 50%, 25%, and 10% of their original resolution (left) and compression ratio measured as bits per second (right). Results are reported on a per frame and per video basis.

because the original videos were of varying resolution, Section 3). At the video level, accuracy remains relatively high for videos at 50% or higher of their original resolution. This corresponds to an average resolution of 892 × 355.

Shown in the right portion of Table 3 are the same frame- and video-level accuracies for videos with progressively lower compression (bits per second) compared to the original video compression. Here, accuracy remains high even for videos at 10% of the original compression rate. This lowest compression corresponds to an average bit rate of 157 kbps.

The loss in accuracy for lower-resolution videos suggests that there are some small-scale features that are important to distinguishing the real from the fake.

5.2 Generalizability

Our frame-to-prompt classifier requires no training, which makes generalizability to new synthesis models more likely. Our two- and multi-class SVMs, however, do require training, and these types of supervised-learning models often struggle to generalize. To evaluate the generalizability of our SVMs, we performed a leave-one-out analysis in which two-class linear SVMs are trained on CLIP embeddings of six of the seven text-to-video models.

Leave-one-out	Frame (%)	Video (%)
BDAnimateDiffLightning	82.3	94.9
CogVideoX5B	82.6	95.7
Lumiere	79.2	91.9
RunwayML	81.4	94.8
StableDiffusion	83.6	96.7
Veo	82.0	95.9
VideoPoet	82.6	96.2

Table 4: Average accuracy of a two-class linear SVM and CLIP embedding, where one text-to-video model is left out of the training set.

Shown in Table 4 are the frame- and video-level accuracies evaluated on just the real and left-out videos. When trained on all seven text-to-video models, accuracy is 96.9% (Table 1). By comparison, accuracy in this leave-one-out condition ranges from a low of 91.9% to a high of 96.7% with an average of 95.2%, showing that our approach generalizes well to previously unseen models.

5.3 Non-Human Motion

Having been trained on only videos containing human motion, we wondered if our two-class SVM would generalize to arbitrary text-to-video models. We evaluated our model on 100 videos generated by Sora [Sora, 2024] and RunwayML [Gen3, 2024] that did not contain any humans or human-motion. With the CLIP embeddings, the model correctly classified 97.1% of these videos as AI-generated showing generalization to non-human motion and to one text-to-video model not seen in training (Sora).

For the frame-to-prompt model, the highest video-level accuracy obtained was for the fine-tuned CLIP embedding and P3 prompt at an accuracy of 97.5%, on par with the human-motion (see Table 2).

This result suggests that our model has not learned something specific to AI-generated human motion but instead has learned something distinct about AI-generated content. From a forensic perspective, this is highly desirable.

We do not yet fully understand what specific properties of the multi-modal embeddings are distinct from AI-generated content. We speculate, however, that because generative-AI models rely on multi-modal embeddings to convert text prompts into images and videos, the extracted embeddings are distinct from those of real content. Another possibility is that generative-AI models may be prompted with distinct properties like level of descriptiveness, yielding more compact content as compared to real videos.

5.4 Talking Heads

We next wondered if our two-class SVM would generalize to a more constrained type of human motion in the form of face-swap and lip-sync deepfakes [Farid, 2022]. We evaluated our model on 100 real videos and 100 deepfake videos from the DeepSpeak Dataset [Barrington *et al.*, 2024]. These videos depict people sitting in front of their webcam responding to questions and prompts from which face-swap and lip-sync deepfakes were created. With the CLIP embeddings, the macro-average model accuracy is 55.8% heavily biased to classifying content as fake at a rate of 93.1% as compared to real at a rate of 15.2%. This is perhaps not surprising since, with the exception of the face, the content in these AI-manipulated videos is authentic.

For the frame-to-prompt model, the highest accuracy of 60.1% is obtained from the fine-tuned CLIP embedding and P2 prompt, significantly lower than for the full-body, human-motion videos (see Table 2).

Here, we do not see generalization with respect to the two-class model or frame-to-prompt. It remains to be seen if both approaches will improve with training in the case of the two-class model, and updating the fine-tuned CLIP with these types of deepfakes.

5.5 CGI

Lastly, we wondered if our two-class SVM would generalize to non-AI-generated video in the form of CGI. We evaluated the trained two-class SVM on videos from the GTA-Human dataset [Cai *et al.*, 2021]. We evaluated our model on 100 GTA-Human videos (cropped around the human movement at a resolution of 640 × 480 pixels). With the CLIP embeddings, the model correctly classified only 39.1% of these videos. Our model does not generalize to CGI.

Result for the frame-to-prompt model were mixed. Averaged across all prompts (P*), accuracy for the CLIP embedding was 76.1%, as compared to 15.0% for JinaCLIP, and 48.9% for FT-CLIP.

Performance for SigLIP across all prompts was perfect at 100%. That is, the SigLIP embedding for these video frames is more similar to the "AI-generated", "unrealistic", or "manipulated" prompts than the "authentic" or "realistic" prompts. We don't fully understand why there is such a large discrepancy here across embedding, but it may be possible that SigLIP was exposed to CGI content during training.

5.6 Comparison to Related Work

Due to the lack of standardized benchmarks, comparing methods remains imperfect. Nevertheless, we provide a comparison to related work by comparing baseline accuracy and generalizability to unseen generative models. The latter is particularly important because true efficacy in the wild will be limited by the ability of detection models to generalize.

When trained on videos from all four text-to-video AI models, the method described in [Vahdati *et al.*, 2024] attains an AUC in the range 98.5 to 99.3. When one model is left out, the AUC drops to between 67.1 and 77.3, revealing relatively poor generalizability.

When trained on videos from all models, our two-class linear SVM with CLIP embeddings attains an accuracy of 96.9%. When one model is left out, the accuracy drops to between 91.9% and 96.7%. Although base-level accuracy is comparable, our method generalizes to videos from unseen models better than earlier approaches.

Although previous approaches have only focused on images, our frame- and video-level accuracy and generalizability are comparable or better than previous approaches [Jia *et al.*, 2024; Cozzolino *et al.*, 2024; Khan and Dang-Nguyen, 2024] (Section 5.6). Our frame-to-prompt is particularly attractive because it requires no explicit learning, making deployment relatively straightforward.

6 Discussion

When we first started to think about the problem of distinguishing real from AI-generated human motion, we thought to take a physics-based approach leveraging advances in 3D human-body modeling [Loper *et al.*, 2023]. We hypothesized that by extracting 3D models of the human body, we could expose implausible dynamics and kinematics in AI-generated motion. We, however, quickly ran into obstacles and found it difficult to consistently and reliably extract 3D models in a wide range of human poses and levels of occlusion.

This led us to take a more learning-based approach. Wanting to avoid exploiting low-level features [Vahdati *et al.*, 2024] vulnerable to laundering, we looked towards more semantic-level features, which, as we have shown, are more resilient to laundering.

As with any authentication scheme, we must consider vulnerability to counter-attack by an adversary. Having already shown resilience to standard laundering, we will eventually need to consider a more sophisticated adversarial attack [Carlini and Farid, 2020].

It has long been the goal – and challenge – of media forensics to extract semantic-level features that can distinguish real from fake or manipulated content. While we cannot say for sure that CLIP embeddings of the form explored here capture truly semantic properties, their resilience to resolution and quality and their generalizability suggest semantic-like representations. This is particularly the case for the surprisingly high performance achieved by our frame-to-prompt approach in which a video-frame embedding is simply compared to a pair of text embeddings of the form "authentic image" and "AI-generated image."

As text-to-video models advance in photorealism and computational efficiency [Yu *et al.*, 2024], complementary models for video-to-video generation and text-based video editing are also improving [Bao *et al.*, 2024]. We should expect that the binary classification of real vs. fake will soon be complicated by hybrid videos that are partially AI-generated or AI-adjusted. Whether semantic methods, such as ours, can capture such subtleties remains to be seen.

Ethical Statement

Unlike the ethical considerations that should be weighed in the creation of generative-AI models (e.g., training data ownership and potential misuse in the form of the creation of non-consensual intimate imagery), we don't anticipate similar concerns with the development of techniques to detect AI-generated content. We recognize, however, that describing a detection technique could further enable the creation of more compelling deepfakes. We believe, however, that the benefit to the scientific field outweighs this risk, and to mitigate this risk we have chosen not to open-source our code.

Acknowledgements

This work was done during the first author's (MB) internship at Google. The first author thanks his manager, Tony Chen, and co-host, Catherine Zhao, for their exceptional support, mentoring, and discussions related to this work; thanks to Prakash Hariramani, Sunita Verma, and Joshy Joseph for their support within leadership; and thanks for helpful discussions to: Amanda Walker, Athula Balachandran, Dara Bahri, David Hendon, Jess Gallegos, John Choi, Karen Guo, Nicholas Carlini, Nick Pezzoti, Stanley Wang, Sven Gowal, and Tom Hume. This work was supported by a gift from YouTube and funding from the University of California Noyce Initiative.

References

Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. Detecting deep-fake videos from appearance and behavior. In *IEEE International Workshop on Information Forensics and Security*, pages 1–6. IEEE, 2020.

Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: A highly consistent, dynamic and skilled text-to-video generator with diffusion models. arXiv:2405.04233, 2024.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. arXiv:2401.12945, 2024.

Sarah Barrington and Hany Farid. People are poorly equipped to detect AI-powered voice clones. arXiv:2410.03791, 2024.

Sarah Barrington, Romit Barua, Gautham Koorma, and Hany Farid. Single and multi-speaker cloned voice detection: From perceptual to learned features. In *IEEE International Workshop on Information Forensics and Security*, pages 1–6. IEEE, 2023.

Sarah Barrington, Matyas Bohacek, and Hany Farid. DeepSpeak dataset v1. 0. arXiv:2408.05366, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv:2311.15127, 2023.

Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, Jessica O'Dell, Kevin Butler, and Patrick Traynor. Who are you (I really wanna know)? Detecting audio DeepFakes through vocal tract reconstruction. In *31st USENIX Security Symposium*, pages 2691–2708, 2022.

Matyáš Boháček and Hany Farid. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proceedings of the National Academy of Sciences*, 119(48):e2216035119, 2022.

Matyas Bohacek and Hany Farid. Lost in translation: Lip-sync deepfake detection from audio-video mismatch. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, Workshop on Media Forensics*, pages 4315–4323, 2024.

Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3D human recovery. arXiv:2110.07588, 2021.

Nicholas Carlini and Hany Farid. Evading deepfake-image detectors with white-and black-box attacks. In *International Conference on Computer Vision and Pattern Recognition Workshop*, pages 658–659, 2020.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. PaLI: A jointly-scaled multilingual language-image model. arXiv:2209.06794, 2022.

John Collomosse and Andy Parsons. To authenticity, and beyond! Building safe and fair generative AI upon the three pillars of provenance. *IEEE Computer Graphics and Applications*, 44(3):82–90, 2024.

Corinna Cortes and Vladimir Naumovich Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of AI-generated image detection with CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4356–4366, 2024.

Soumyya Kanti Datta, Shan Jia, and Siwei Lyu. Exposing lip-syncing deepfakes from mouth inconsistencies. arXiv:2401.10113, 2024.

Vincenzo De Rosa, Fabrizio Guillaro, Giovanni Poggi, Davide Cozzolino, and Luisa Verdoliva. Exploring the adversarial robustness of CLIP for AI-generated image detection. arXiv:2407.19553, 2024.

Henrique Galvan Debarba, Sylvain Chagué, and Caecilia Charbonnier. On the plausibility of virtual body animation features in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(4):1880–1893, 2020.

Alexander Diel, Sarah Weigelt, and Karl F Macdorman. A meta-analysis of the uncanny valley's independent and dependent variables. *ACM Transactions on Human-Robot Interaction*, 11(1):1–33, 2021.

Hany Farid. Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety*, 1(4), 2022.

RunwayML Gen3. https://runwayml.com/research/introducing-gen-3-alpha, 2024.

Shan Jia, Xin Li, and Siwei Lyu. Model attribution of face-swap deepfake videos. In *IEEE International Conference on Image Processing*, pages 2356–2360. IEEE, 2022.

Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. Can ChatGPT detect deepfakes? A study of using multimodal large language models for media forensics. In *International Conference on Computer Vision and Pattern Recognition Workshop*, pages 4324–4333, 2024.

Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. CLIPping the deception: Adapting vision-language models for universal deepfake detection. In *International Conference on Multimedia Retrieval*, pages 1006–1015, 2024.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. VideoPoet: A large language model for zero-shot video generation. arXiv:2312.14125, 2023.

Andreas Koukounas, Georgios Mastrapas, Michael Günther, Bo Wang, Scott Martens, Isabelle Mohr, Saba Sturua, Mohammad Kalim Akram, Joan Fontanals Martínez, Saahil Ognawala, et al. Jina CLIP: Your CLIP model is also your text retriever. arXiv:2405.20204, 2024.

Shanchuan Lin and Xiao Yang. AnimateDiff-Lightning: Cross-model diffusion distillation. arXiv:2403.12706, 2024.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

Rachel McDonnell, Martin Breidt, and Heinrich H Bülthoff. Render me real? Investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics*, 31(4):1–11, 2012.

Peter Neri, M Concetta Morrone, and David C Burr. Seeing biological motion. *Nature*, 395(6705):894–896, 1998.

Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1001–1010, 2024.

Sophie J Nightingale and Hany Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, 2022.

Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *IEEE/CVF International Conference on Computer Vision*, pages 7184–7193, 2019.

Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6111–6121, 2021.

Pexels. https://www.pexels.com, 2024.

Alessandro Pianese, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Deepfake audio detection by speaker verification. In *IEEE International Workshop on Information Forensics and Security*, pages 1–6. IEEE, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of clip-filtered 400 million image-text pairs. arXiv:2111.02114, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can CLIP benefit vision-and-language tasks? arXiv:2107.06383, 2021.

Stefan Smeu, Elisabeta Oneata, and Dan Oneata. DeCLIP: Decoding CLIP representations for deepfake localization. arXiv:2409.08849, 2024.

Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. CLIP models are few-shot learners: Empirical studies on vqa and visual entailment. arXiv:2203.07190, 2022.

Sora. https://openai.com/index/sora/, 2024.

Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: Learning lip sync from audio. *ACM Transactions on Graphics*, 2017.

Danial Samadi Vahdati, Tai D Nguyen, Aref Azizpour, and Matthew C Stamm. Beyond deepfake images: Detecting AI-generated videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, Workshop on Media Forensics*, pages 4397–4408, 2024.

Veo. https://deepmind.google/technologies/veo, 2024.

Sviatoslav Voloshynovskiy, Shelby Pereira, Thierry Pun, Joachim J Eggers, and Jonathan K Su. Attacks on digital watermarks: Classification, estimation based attacks, and benchmarks. *IEEE Communications Magazine*, 39(8):118–126, 2001.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. arXiv:2408.06072, 2024.

Heng Yu, Chaoyang Wang, Peiye Zhuang, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Laszlo A Jeni, Sergey Tulyakov, and Hsin-Ying Lee. 4Real: Towards photorealistic 4D scene generation via video diffusion models. arXiv:2406.07472, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

A Experimental Setup

A.1 CLIP

The following model was used:
https://huggingface.co/openai/clip-vit-base-patch32

A.2 SigLIP

The following model was used:
https://huggingface.co/google/siglip-base-patch16-224

A.3 JinaCLIP

The following model was used:
https://huggingface.co/jinaai/jina-clip-v1

A.4 FT-CLIP

The model was fine-tuned for one warm-up epoch at a learning rate of 10^{-5} and one regular epoch at a learning rate of $2 \cdot 10^{-3}$. An SGD optimizer with a cosine learning rate scheduler was employed for both epochs. The batch size was set to 16 for training and 100 for testing. The training set comprised two balanced sets of real and AI-generated frames, with captions constructed as *"a {REAL/FAKE} image of {ACTION PROMPT}"*. Training was performed on an A100 GPU for approximately eight hours.

A Video Generation Prompts

A person walking through a park.	A person dancing in their living room.
A person doing yoga in their backyard.	A person cooking a meal in the kitchen.
A person playing fetch with their dog.	A person riding a bicycle down the street.
A person jogging on a sidewalk.	A person lifting weights at home.
A person practicing a musical instrument.	A person painting on a canvas.
A person doing push-ups in their living room.	A person stretching before a workout.
A person reading a book in a cozy chair.	A person writing in a journal.
A person meditating in a quiet room.	A person gardening in their backyard.
A person playing a board game with family.	A person folding laundry.
A person making a bed.	A person doing a puzzle.
A person brushing their hair.	A person tying their shoes.
A person washing dishes.	A person vacuuming the living room.
A person watering plants.	A person sewing on a sewing machine.
A person knitting on the couch.	A person ironing clothes.
A person mopping the floor.	A person baking cookies.
A person eating a meal at the table.	A person talking on the phone.
A person brushing their teeth.	A person playing with a pet.
A person working on a laptop.	A person watching TV.
A person drinking coffee on the porch.	A person taking a selfie.
A person organizing a bookshelf.	A person practicing calligraphy.
A person doing sit-ups.	A person practicing a dance move in front of a mirror.
A person applying makeup.	A person trimming a plant.
A person playing catch in the backyard.	A person skating in a driveway.
A person raking leaves.	A person cleaning windows.
A person decorating a cake.	A person unboxing a package.
A person eating popcorn while watching a movie.	A person putting together a DIY project.
A person reading a bedtime story.	A person practicing a speech.
A person blowing out birthday candles.	A person organizing their closet.
A person playing an online game.	A person having a video call.
A person building a model.	A person practicing origami.
A person doing jumping jacks.	A person dancing to music.
A person coloring in a coloring book.	A person taking a walk with a friend.
A person writing a letter.	A person enjoying a picnic.
A person birdwatching in the backyard.	A person making a smoothie.
A person cutting paper for a craft.	A person playing with building blocks.
A person wrapping a gift.	A person lighting a candle.
A person drawing a picture.	A person setting the table.
A person playing with a toy.	A person cleaning a mirror.
A person arranging flowers.	A person holding a pet.
A person organizing a drawer.	A person folding a paper airplane.
A person playing a card game.	A person practicing a yoga pose.
A person writing a grocery list.	A person doing a handstand against a wall.
A person making a sandcastle.	A person playing a sport in the backyard.
A person performing a simple magic trick.	A person showing a thumbs-up.
A person waving hello.	A person doing a dance move.
A person practicing a hobby.	A person clapping their hands.
A person jumping with joy.	A person making a funny face.
A person giving a high-five.	A person giving a thumbs-down.
A person walking up stairs.	A person walking down the stairs.
A person walking down a street.	A person jogging on a track.