From *Lifestyle Vlogs* to Everyday Actions
Supplemental Material

David F. Fouhey, Wei-cheng Kuo, Alexei A. Efros, Jitendra Malik
EECS Department, UC Berkeley

Abstract

Overview

Thank you for reading our supplemental material. We have put information about VLOG here that you may find useful. Each section is numbered according to the section in the paper it refers to, and to avoid having to flip through image collections, text-based things come first.

- **Supplemental Video Details**: this describes the content of the supplemental video.
- §**3 – Acquisition Pipeline Details**: this describes the pipeline that finds our dataset.
- §**4 – Additional statistics**: this provides additional data statistics about VLOG.
- §**5 –Long-tail items**: this provides the list of items found in our sample, which demonstrates the long tail of items naturally present in VLOG.
- §**8 – Experimental details**: this provides some details about the method we use to predict future hand locations.
- **Samples for** §**5 – Per Category Items**: this provides positive samples for each category of hand/semantic object interaction.
- **Samples for** §**6 – Scene Samples**: this provides a comparison of the scenes of VLOG (e.g., bedroom/bathroom/etc) as depicted in a scene classification dataset vs. as depicted in VLOG.

Supplemental Video Details

VLOG and similar efforts are video-based and **we highly encourage watching the supplemental video**.

1. The first part shows the data (the first 10s of clips) for VLOG in comparison to other similar data gathering efforts (Watch-n-Patch, CAD-120, Charades). Even looking at a handful of videos from in-house efforts reveals the difficulty of scaling up in terms of diversity of person, action, and environment. Charades was a large step forward, but VLOG's underlying data is at least as good as Charades while being vastly more plentiful, as shown by the bar charts. This is true for both categories that Charades gathers (e.g., refrigerator) and especially so for ones it did not (e.g., microwave).

2. Finally, the video shows random samples. Even if one views the videos in this montaged format (watching 8×10 videos simultaneously), watching the entirety of VLOG would take 4.3 hours.

§3 – Acquisition Pipeline Details

This is somewhat involved and difficult, but primarily an engineering, not research task. It is documented here to answer technical questions.

Finding Videos. Youtube provides, as an undocumented feature, four thumbnails. We download these and extract Alexnet `pool5` features on these thumbnails. Our feature vector is the average `pool5` feature plus the min, mean, and max distance between the activations. These distance features help find videos where someone is talking to the camera the entire time.

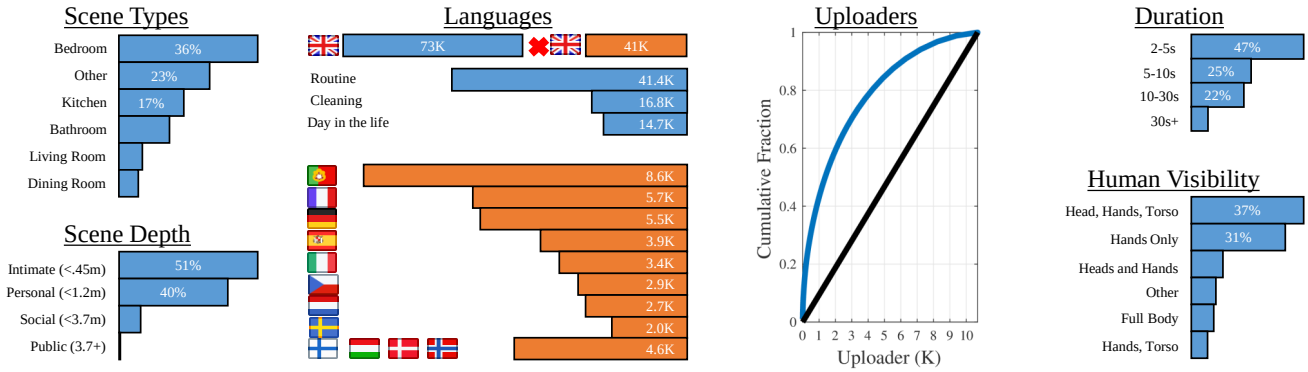
Finding Episodes within Videos. We use an approach based on SIFT homography fitting after experimenting with a number of alternate approaches. Our method scans every 10 frames, looking for unmatchable frames and large movements. Since we are especially interested in static clips and many clips are already static, we first look for support for the identity transformation, and only then start RANSAC iterations. This saves considerable computation.

1. We first run shot detection every 10 frames ($\sim$ 3Hz)
2. We then scan forward every 30, 60, and 90 frames (where the clip is long enough for this), and verify that there is some evidence of a match. Frames that cannot be matched forward tend to be during dissolves, which are frequent enough to require removal. We mark a discontinuity at any point where the frame cannot be matched.
3. Finally, given the shot boundaries at every 10 frames, we examine the 10 frames in the middle. This fine-scale detection is crucial: many videos are aggressively trimmed and not doing this results in premature cuts.

For N frames that are segmented into k shots, this procedure requires $N/10 + 10 * (k - 1)$ SIFT feature extractions as opposed to N, and similar savings on homography fits.

Finding Clips. We re-run the video classifier; here, since the clips typically have very similar interframe appearance, we retrain the classifier using only the average CNN activation feature, and not the inter-frame distance features.

§4 – Additional statistics



Here, we provide additional statistics about VLOG: where it comes from, the distribution of types (scene depth and class), a CDF of uploaders, video lengths (average length $\approx$ 10s), and the visibility of various body types.

§5 – Long-tail items

We provide the name (post canonicalization) of the objects that humans touch and their count in our 500 image sampling of the dataset. We note that many of the objects are probably more naturally referred to more specifically (e.g, "shampoo" instead of "bottle", "crackers" instead of "box"). Note the large number of common, natural categories that simply appear once: plant, zucchini, bathtub, etc.

Categories: bottle (36), cellphone (31), makeupbrush (16), blanket (14), laptop (14), spoon (13), bag (12), body (12), jar (11), notebook (10), tube (10), book (9), box (9), makeupcompact (9), shirt (9), bowl (8), knife (8), paper (8), table (8), toothbrush (7), pillow (6), cabinet (5), cup (5), drawer (5), fork (5), glas (5), hairstraightener (5), lipstick (5), pan (5), shoe (5), towel (5), backpack (4), bread (4), doll (4), door (4), facewipe (4), hairbrush (4), mug (4), banana (3), bed (3), carton (3), meat (3), mop (3), nailpolish (3), plate (3), refrigerator (3), sheet (3), sink (3), sponge (3), sweatshirt (3), toothpaste (3), cat (2), chair (2), cheese (2), counter (2), dish (2), dog (2), facebrush (2), floor (2), hairdryer (2), hanger (2), iron (2), monitor

(2), pen (2), pitcher (2), pot (2), remote (2), spraybottle (2), stuffedanimal (2), toy (2), yogamat (2), apple (1), armwarmer (1), baby (1), babyjumper (1), basket (1), bathtub (1), blender (1), bookshelf (1), butter (1), cage (1), calculator (1), car (1), cheesegrater (1), coffeetable (1), container (1), cookie (1), diaper (1), dirt (1), dishwasher (1), dresser (1), drill (1), drumstick (1), dumbbell (1), duster (1), egg (1), espressomachine (1), facialmask (1), flower (1), folder (1), gamecontroller (1), icecubetray (1), jewelryholder (1), kettle (1), lamp (1), lettuce (1), lid (1), magazine (1), measuringcup (1), muffin (1), napkin (1), nightstand (1), pacifier (1), paintbrush (1), paintroller (1), pencil (1), piano (1), pie (1), plant (1), popsiclemold (1), powercord (1), purse (1), rag (1), sand (1), sander (1), sewingmachine (1), shelf (1), skirt (1), stove (1), straw (1), tape (1), tin (1), tray (1), wardrobe (1), watch (1), zucchini (1) .

§8 – Experimental details

Hand prediction Let: $C(k, s)$ denote a convolution of k kernels with size $s \times s$; R, BN denote ReLU and Batchnorm.

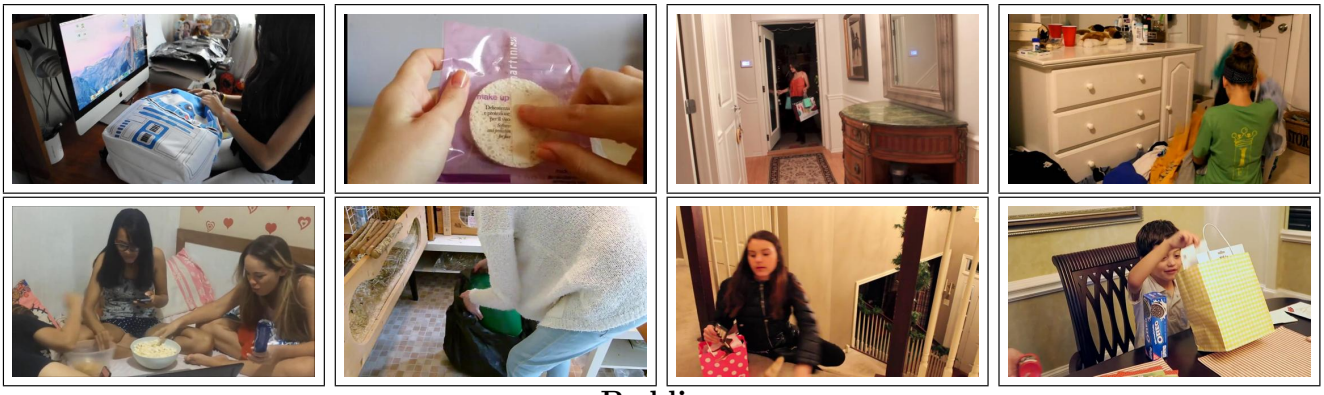
1. The image I is passed through the base DRN-D-54 network ϕ, yielding a 512 channel feature map.
2. The time offset variable δ is upsampled to feature map size, then mapped to $64D$ through two layers ($C(16, 1) \rightarrow R \rightarrow C(64, 1) \rightarrow R$), or in total ψ.
3. After concatenating image $\phi(I)$ and time features $\psi(\delta)$, we predict the final output passing it through three 3×3 convolutions (denoted ζ in total): $C(128, 3) \rightarrow BN \rightarrow R \rightarrow C(128, 3) \rightarrow BN \rightarrow R \rightarrow C(2)$, followed by $8\times$ bilinear upsampling.

In total the network is $\zeta(\text{cat}(\phi(I), \psi(\delta)))$ The method is trained to minimize a cross-entropy loss.

Samples for §5 – Per Category Items

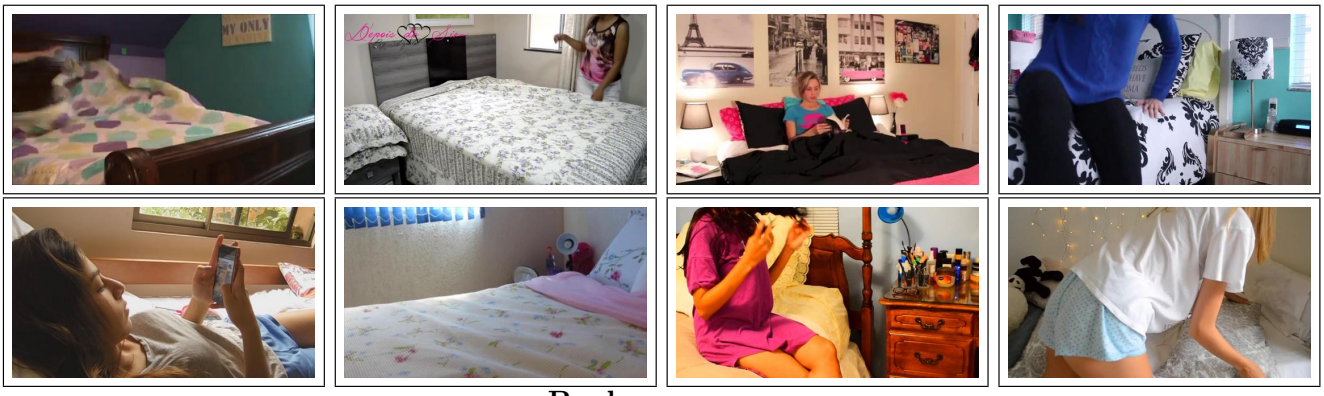
We now show samples from each of our 30 categories. Note that these are middle frames from the clips, so the person may or many not be interacting with the object at the time. Best viewed with zoom.

Bag

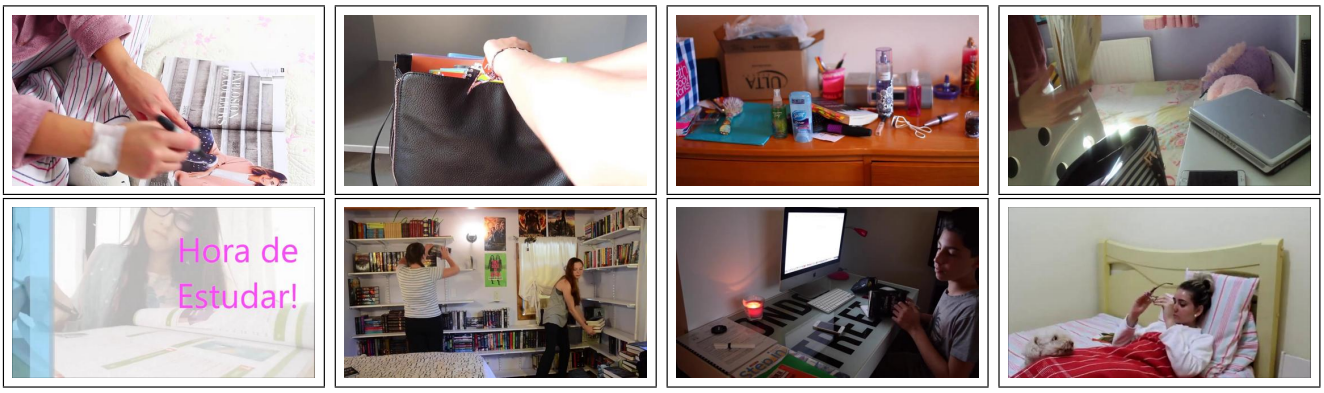


Bedding

Bed



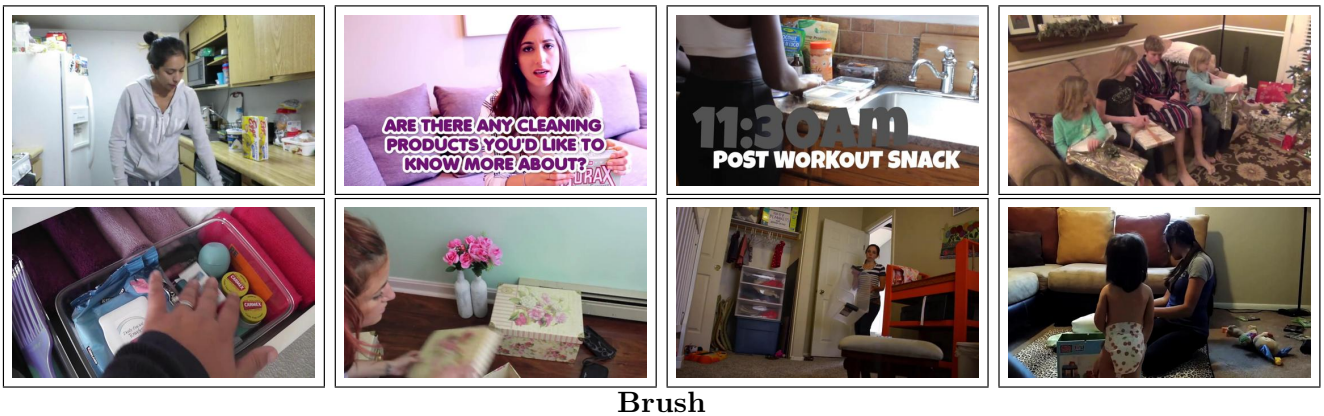
Book or papers



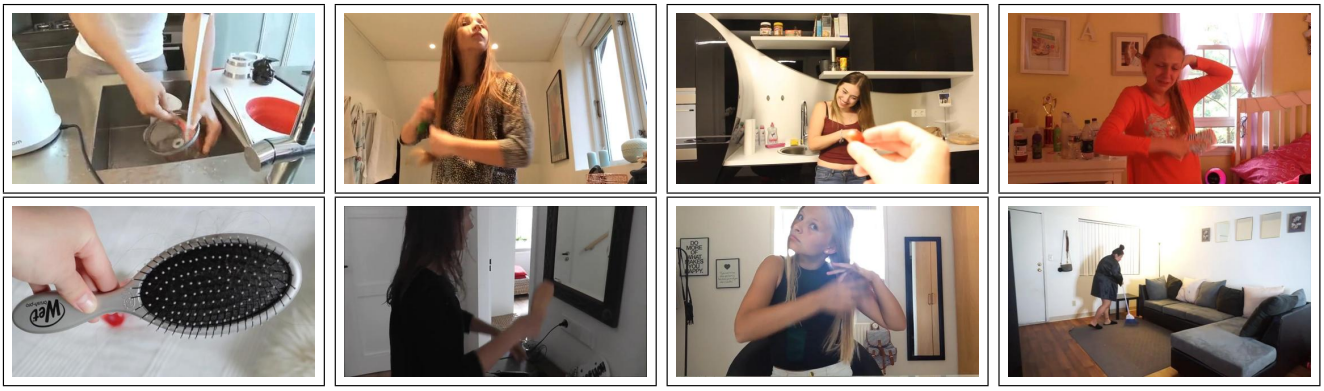
Bottle/tube

Bowl

Box

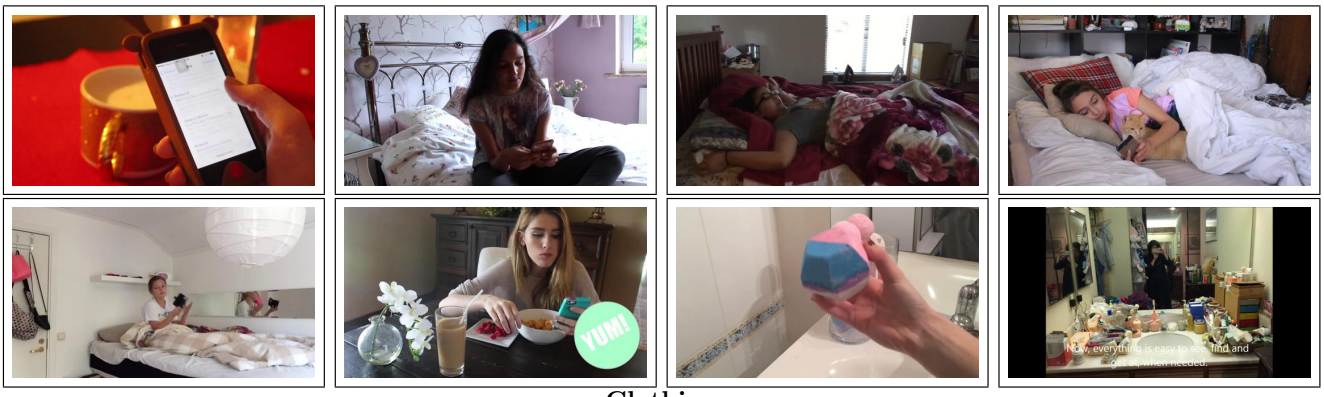


Brush

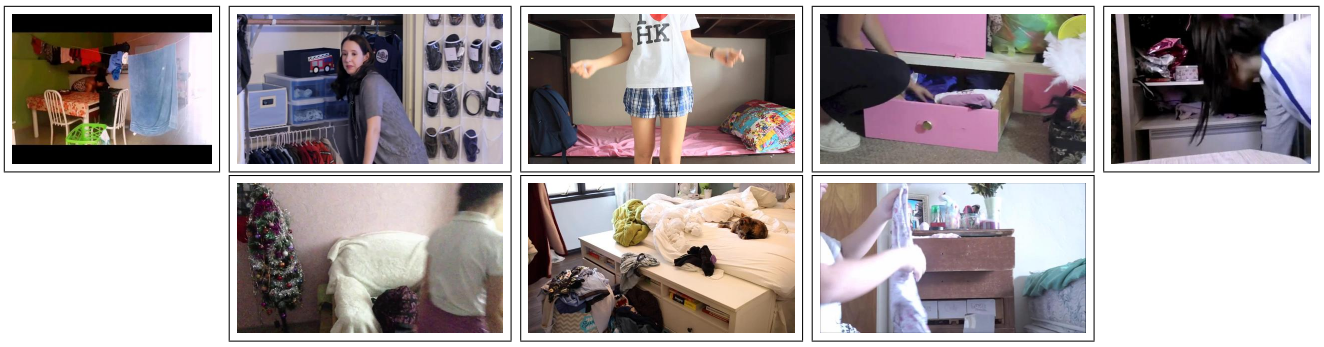


Cabinet

Cell phone

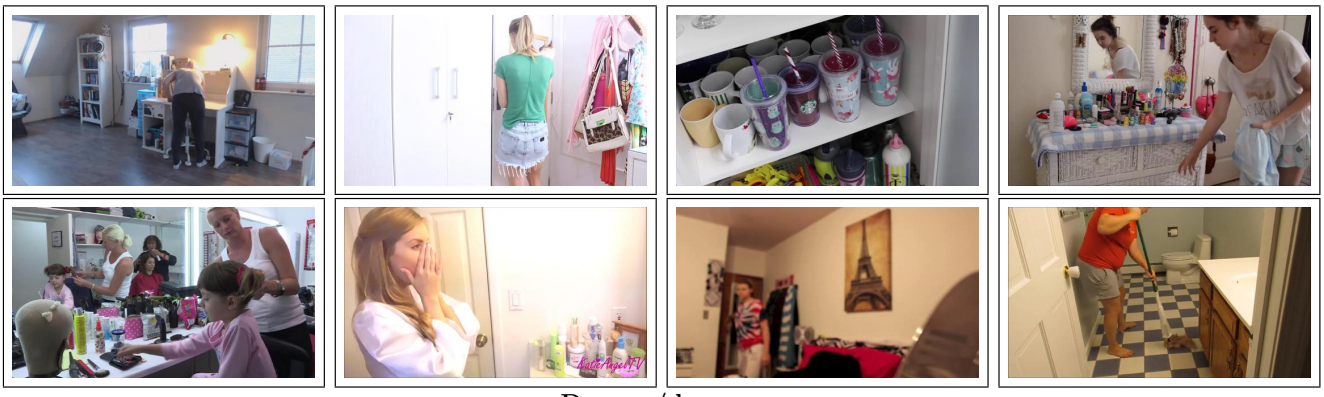


Clothing

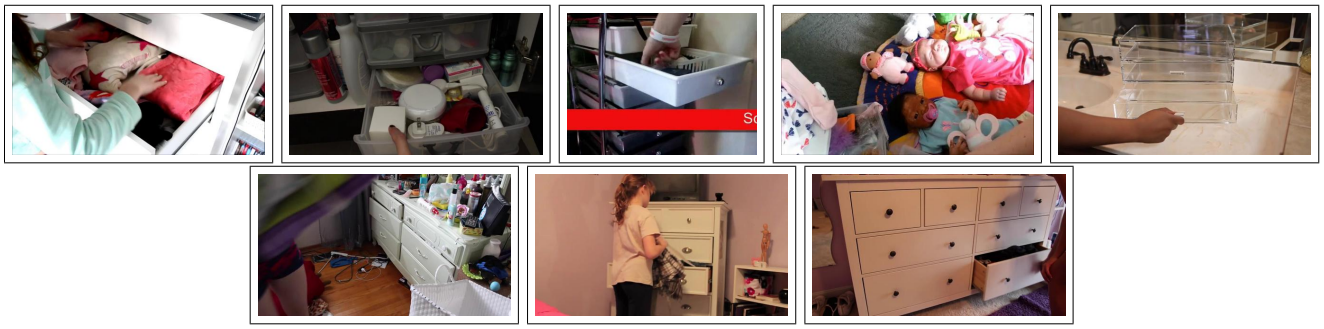


Cup

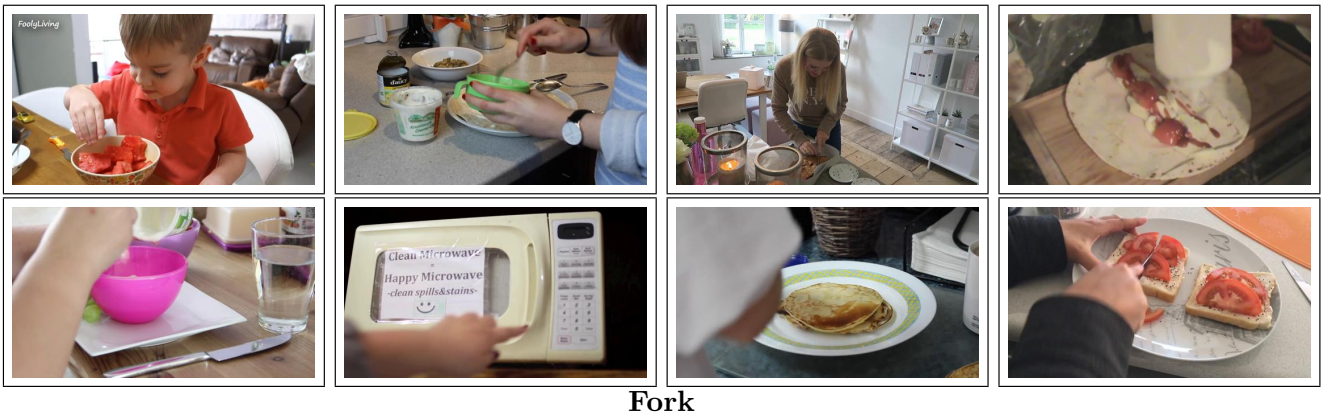
Door



Dresser/drawers



Food



Fork

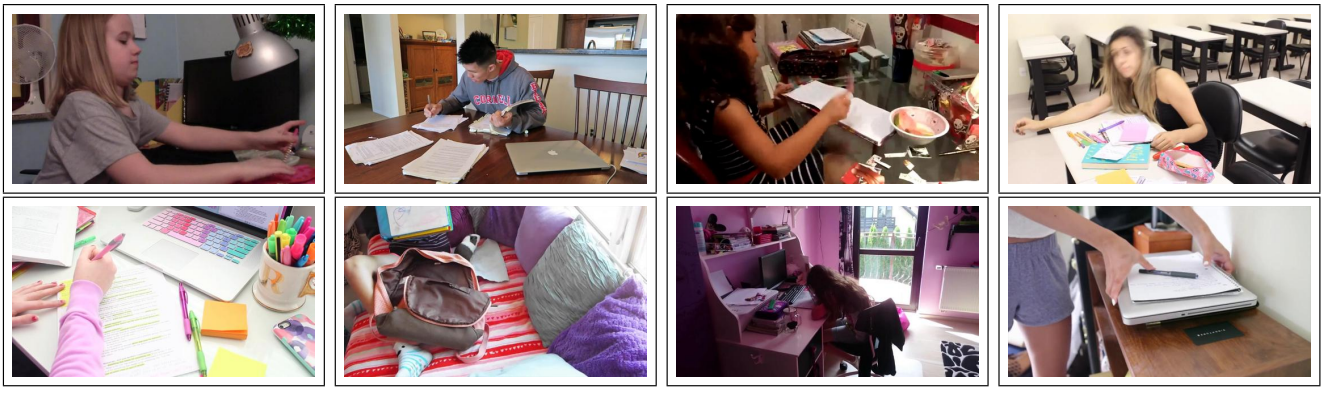
Knife

Laptop

Microwave

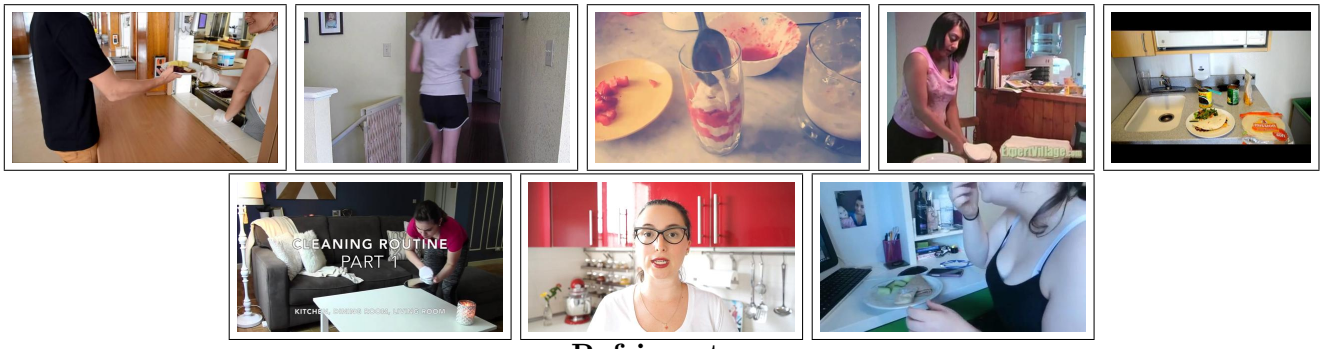
Oven

Pen or pencil



Pillow

Plate

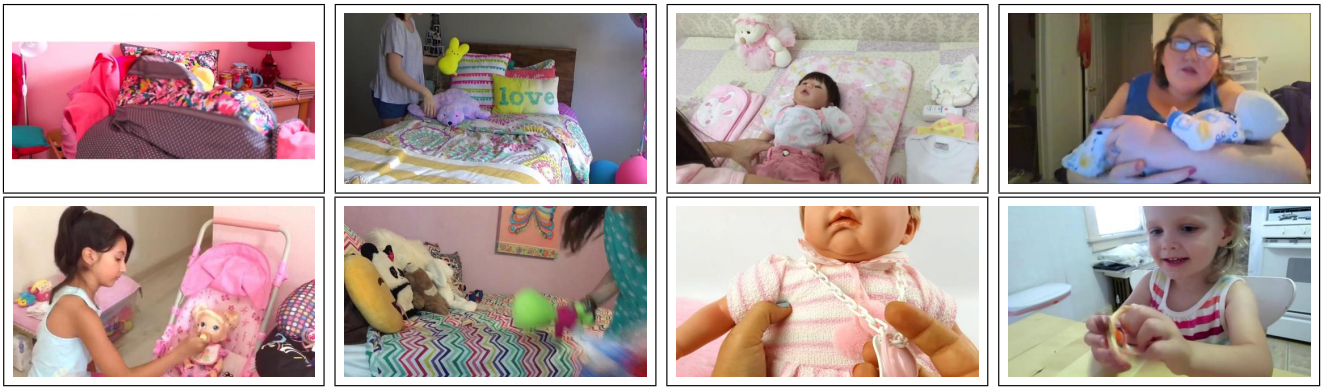


Refrigerator

Sink

Spoon

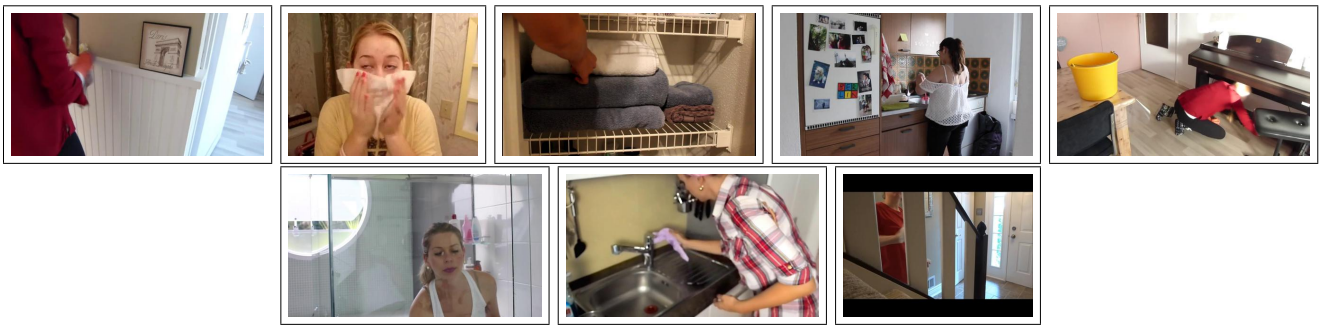
Stuffed animal or doll



Table

Toothbrush

Towel



Samples for §6 – Scene Samples

We now show samples from both Places365 and VLOG for the five scene categories of VLOG. This is best viewed with zoom. Note the overall sterility and frequency of "real estate photography" camera angles in Places365. In bathrooms and kitchens, note the lack of clutter; and in bedrooms, the fact that every single bed is made.

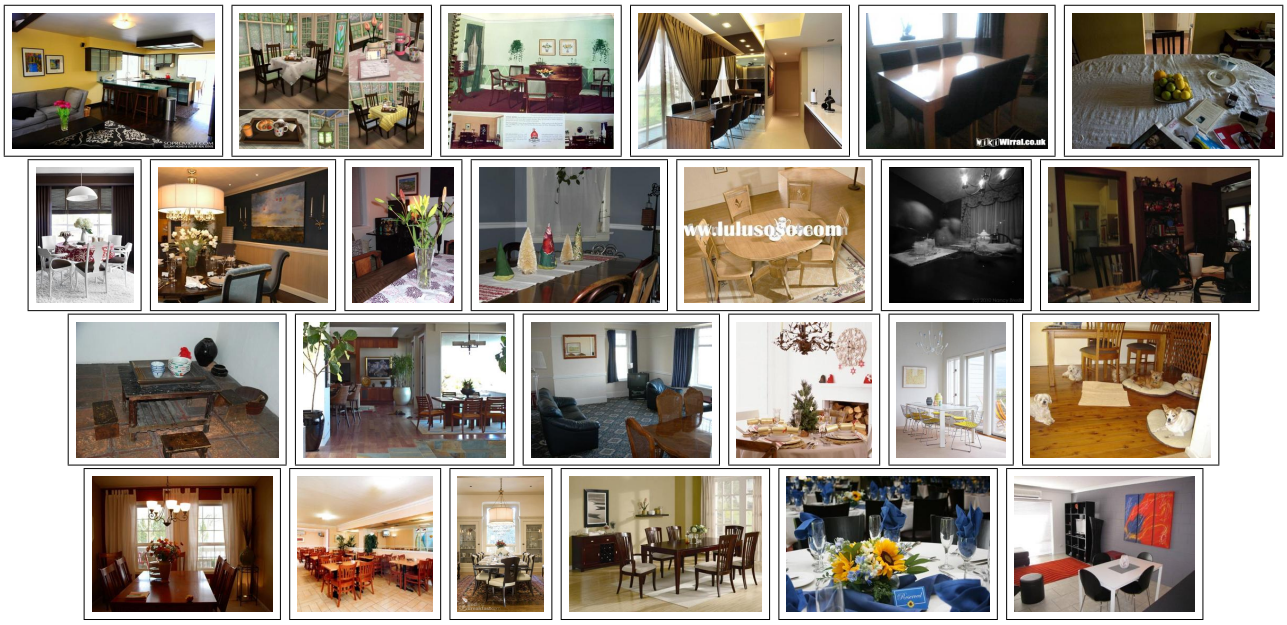
Places365 Bathroom

VLOG Bathroom

Places365 Bedroom

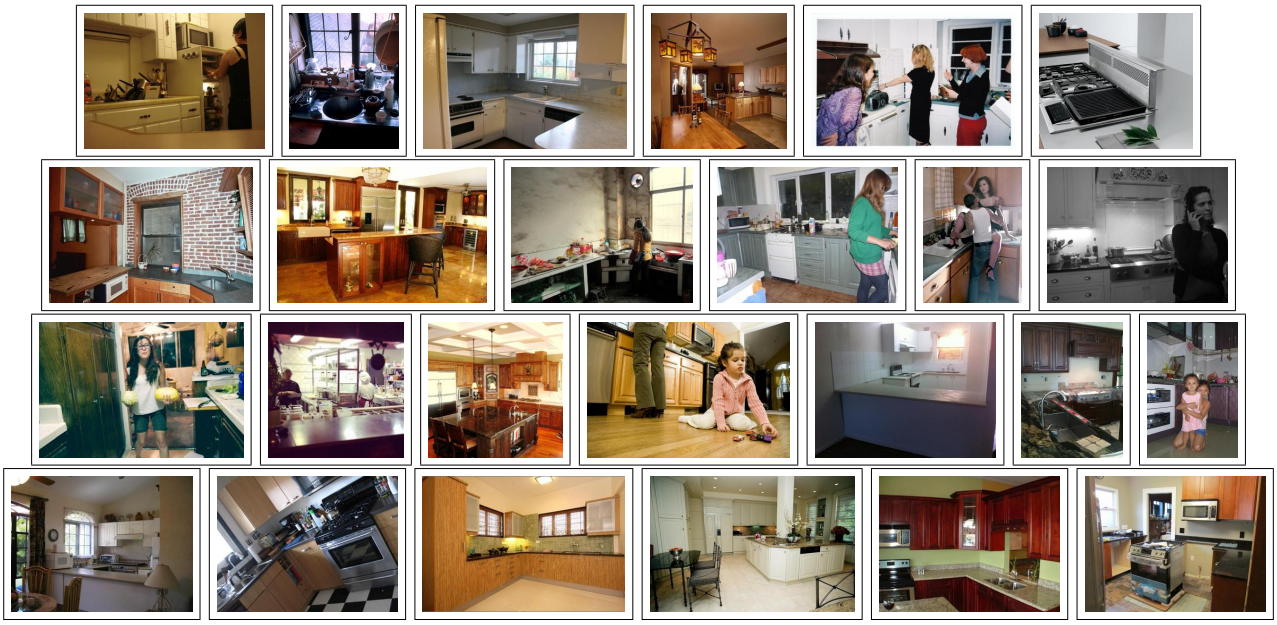
VLOG Bedroom

Places365 Dining Room



VLOG Dining Room

Places365 Kitchen



VLOG Kitchen

Places365 Living Room

VLOG Living Room