# Look Over Here: Attention-Directing Composition of Manga Elements

*Supplementary Material*

Ying Cao Rynson W.H. Lau Antoni B. Chan

Department of Computer Science, City University of Kong Kong

## 1 Data Acquisition and Preprocessing

To train our probabilistic model, we have collected a data set comprising 80 manga pages from three chapters of three different series: "*Bakuman*", "*Vampire Knight*" and "*Fruit Basket*". These manga series have distinctive composition complexities and patterns, in order for our dataset to be able to capture a wide range of composition styles used by manga artists.

**Annotation.** We manually segmented and annotated all the pages in our dataset. Each page was segmented into a set of panels. For each panel, we first manually label its shot type (i.e., long, medium, close-up, or big close-up) and segmented the foreground subjects and their corresponding balloons. Given a set of segmented panels across all the pages, we then partitioned the panels into three groups with similar geometric features, which include aspect ratio (i.e., width/height) and size (i.e., area). The clustering was done using a Gaussian mixture model (GMM), initialized by k-means clustering. Grouping geometrically similar panels allows the probabilistic model to learn composition patterns that vary with the panel shape.

**Eye-tracking Data from Multiple Viewers.** To understand how manga artists control viewer attention via composition of subjects and balloons, we have conducted a visual perception study to track the participants' eye movements as they read the manga pages in our dataset.

Thirty participants with various background were recruited from a university participant pool. We required the participants to have some experience reading manga. Each participant was asked to continuously view around 30 manga pages from one chapter of one of our three manga series, and was told that they would be asked several questions at the end of the viewing session. The questions are based on comprehension and in the form of multiple choices. Participants who gave wrong answers to half of the questions were excluded from further consideration. Each session usually lasted for about 15 minutes. The same participant could join more than one session but was not allowed to view the same series twice.

For eye-tracking, we used the Eyelink 1000 system. The participants sat in front of the computer monitor, with their heads fixed on a chin-rest device and at a distance of $70cm$ from the screen. Eye fixations were recorded at a sampling rate of $250Hz$.

At the end of the study, we had eye movement data from 10 different participants for every page in our dataset. The saccades (i.e., the rapid eye movements between eye fixations) in the eye movement data indicate how the viewers transition their attention between the panel elements (i.e., subjects and balloon) of interest. To compactly and visually represent such information, we preprocess the raw eye movement data to build an *element graph.* In the graph, each node represents a panel element, and each edge represents a transition of viewer attention between two elements. The thickness of the edge is proportional to the number of viewers following that route. Note that a transition can be bi-directional because the viewers might read back and forth to explore contents of interest.

In this stage, we obtain a set of training examples $\mathcal{D} = \{\mathcal{P}_i\}$. Let $\mathcal{P}_i = \{\{T_j\}, \{\mathbf{V}_j\}, \{\mathcal{S}^j_k\}, \{\mathcal{B}^j_k\}, \mathcal{G}_i\}$ be the $i^{th}$ page. $T_j$ and $\mathbf{V}_j$ are the shot type and the geometric configuration of the $j^{th}$ panel, including its center location and geometric cluster index (obtained by panel clustering in the annotation step). $\mathcal{S}^j_k$ and $\mathcal{B}^j_k$ denote the $k^{th}$ subject and balloon in the $j^{th}$ panel, respectively. Each element is represented as $(\mathbf{x}, r)$, where $\mathbf{x}$ is the center of its bounding box and $r$ is its size computed as square root of the product of the bounding box's width and height. $\mathcal{G}_i$ is a binary matrix storing if there is a viewer attention transition between a pair of elements on the $j^{th}$ page. Attention transition from one element to another is thought of as being present, only if more than 50% of the viewers transition through the route.

## 2 Constraint Terms in the Likelihood

The formulation about *overlap constraint term*, *order constraint term*, and *subject relation constraint term* are detailed as follows.

- The *overlap constraint term* ($C_{\text{overlap}}$) is defined as

$$C_{overlap} = \sum_{p} \sum_{(e_i,e_j)\in E^p} \left(1 - \frac{\mathcal{A}(e_i \cap e_j)}{\min[\mathcal{A}(e_i), \mathcal{A}(e_j)]}\right), \tag{1}$$

  where $E^p$ is the set of elements in panel $p$, and $\mathcal{A}(\cdot)$ is the area of a polygon.

- The *order constraint term* ($C_{\text{order}}$) penalizes the configurations that violate the reading order among a sequence of balloons. We denote $B^p$ as the set of balloons in panel $p$ and $RO(b_i)$ as the desired reading order of balloon $b_i$. The set of balloons that should be read *after* $b_i$ in panel $p$ is $B^p_i = \{b_j | RO(b_i) < RO(b_j)\}$. The order constraint term is

$$C_{\text{order}} = \sum_{p} \sum_{b_i \in B^p} \frac{1}{|B^p_i|} \sum_{b_j \in B^p_i} \psi(b_i, b_j), \tag{2}$$

  where $\psi(b_i, b_j)$ is 1 if $b_i$ and $b_j$ are in correct order, and 0 otherwise. To determine if balloons $i$ and $j$ are arranged in correct order, we use the representation described in [CRHC06]. Specifically, as illustrated in Figure 1, we construct an occupancy region (the shaded area) for balloon $i$. Given $RO(b_i) < RO(b_j)$, we define balloons $i$ and $j$ as being in the correct order, only when the center of balloon $j$ is located outside of the occupancy region, i.e., the green balloon in Figure 1.

$RO(i) < RO(j)$

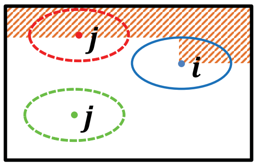


Figure 1: Determining if balloons $i$ and $j$ are in correct order.

- The *subject relation constraint term* ($C_{\text{relative}}$) is defined as:

$$C_{\text{relative}} = 1 - \frac{\|\mathbf{v}_{e,t} - \mathbf{v}_u\|}{2\max(\|\mathbf{v}_{e,t}\|, \|\mathbf{v}_u\|)}, \tag{3}$$

  where $\mathbf{v}_u$ is the relative position resulting from the composition, and $\mathbf{v}_{e,t}$ is a semantic-specific vector, representing most likely relative position for $e$ and $t$. To estimate $\mathbf{v}_{e,t}$, we build a probability table $P(e,t)$, with each entry storing a probability distribution of relative vectors for a joint configuration of $e$ and $t$ from our dataset. The probability distributions are obtained by identifying all pairs of interacting subjects in our dataset and fitting a bivariate Gaussian function to the pair-wise vectors. Given $e$ and $t$, $\mathbf{v}_{e,t}$ is generated by sampling the proper probability distribution in the table.

## 3 Parameterization of the Likelihood of Attention Transition ($E_{pair}$)

Formally, $E_{pair} = w_1 E_I + w_2 E_D + w_3 E_O + w_4 E_S$, where $\{w_k\}$ are weights that balance the contributions of each term, and are also parameters of the sigmoid function. The *identity term* $E_I$ encourages an attention transition to happen between the same types of elements. It is defined as $\delta(I_i, I_j)$, where $I_i \in \{\text{subject}, \text{balloon}\}$ and $\delta(\cdot)$ is the Kronecker delta function that is 1 when the variables are equal, and 0 otherwise. The *distance term* $E_D$ promotes attention transition between elements that are close in spatial distance, and is defined as $\frac{d_{ij}}{L}$, where $d_{ij}$ is the distance between $i$ and $j$, while $L$ is the diagonal length of the panel.

The *orientation term* $E_O$ encourages an attention transition when $j$ is below and to the left of $i$. As illustrated in Figure 2, when $j$ falls within the shaded region of the local coordinate system of $i$, $E_O$ is set to 1, and 0 otherwise. Since the reading order for manga is from top to down, and then right to left, the viewer is more likely to move from $i$ to $j$ when they are in this relative configuration. The *scale term* $E_S$ is defined as the ratio between the size of $j$ and that of $i$, and favors the case when the viewer moves from a smaller element to a larger one. $E_{context}$ is designed to contribute negatively to the potential function, which reduces the probability of attending from $i$ to $j$ if $j$ has any strong competitors in its neighborhood. For example, if there is an element $k$ that is closer to $i$ than $j$, it should have a higher probability that the viewer shifts attention from $i$ to $k$ rather than to $j$. It is formally written as $-\frac{1}{|\mathbf{N}_{ij}|} \sum_{k \in \mathbf{N}_{ij}} E_{pair}(\mathbf{o}_i, \mathbf{o}_k)$.

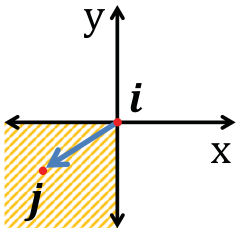


Figure 2: Relative orientation of $j$ to $i$ that encourages attention transitions.

## 4 Estimating the Parameters of f's CPDs in the EM Algorithm

We consider the problem of finding maximum likelihood estimates of the parameters of $\mathbf{f}$'s CPDs. We focus our discussion on parameter estimation of $\mathbf{x^f}$ in the EM algorithm. The same method can also be applied to $\mathbf{y^f}$. For brevity, we drop the superscript $\mathbf{f}$ of $\mathbf{x^f}$ in the derivation below.

Treating $\mathbf{x}$ as a subset of a Gaussian process $x(t) \sim \mathcal{GP}(m(t), k(t, t^{'}))$ gives:

$$\begin{aligned} P(\mathbf{x}) &= \mathcal{N}(\mathbf{0}, \mathbf{K}_{\boldsymbol{\theta}}) \\ &= (2\pi)^{-\frac{n}{2}} |\mathbf{K}_{\boldsymbol{\theta}}|^{-\frac{1}{2}} \exp\left(\mathbf{x}^T \cdot \mathbf{K}_{\boldsymbol{\theta}}^{-1} \cdot \mathbf{x}\right), \end{aligned} \tag{4}$$

where $\boldsymbol{\theta}$ is a hyperparamter of the Gaussian process to be estimated. Note that we assume $m(t) = 0$ as we can always shift the data to accommodate the mean. Since $\mathbf{x}$ has no parents in the probabilistic model, assuming we have a set of i.i.d training examples $\{\mathbf{e}_i\}_{i=1}^T$, complete-data log likelihood is written as:

$$L(\boldsymbol{\theta}; \{\mathbf{x}_i\}) = \sum_i^T \log P(\mathbf{x}_i). \tag{5}$$

**E-step**: We compute conditional expectation of $L$ over the unobserved random variables of $\{\mathbf{x}_i\}$ given evidence $\{\mathbf{e}_i\}$ under current estimate $\boldsymbol{\theta}^{(t)}$ of $\boldsymbol{\theta}$. Dropping the term that is independent of $\boldsymbol{\theta}$, $\{\mathbf{x}_i\}$ yields:

$$E_{\boldsymbol{\theta}^{(t)}}[L(\boldsymbol{\theta}; \{\mathbf{x}_i\}) | \{\mathbf{e}_i\}] = -\frac{T}{2} \log |\mathbf{K}_{\boldsymbol{\theta}}| - \frac{1}{2} \sum_i E_{\boldsymbol{\theta}^{(t)}}[\mathbf{x}_i^T \cdot \mathbf{K}_{\boldsymbol{\theta}}^{-1} \cdot \mathbf{x}_i | \mathbf{e}_i]. \tag{6}$$

As $E_{\boldsymbol{\theta}^{(t)}}[\mathbf{x}_i^T \cdot \mathbf{K}_{\boldsymbol{\theta}}^{-1} \cdot \mathbf{x}_i | \mathbf{e}_i]$ in Equation 6 cannot be computed analytically, we approximate it using Monte Carlo integration:

$$E_{\boldsymbol{\theta}^{(t)}}[\mathbf{x}_t^T \cdot \mathbf{K}_{\boldsymbol{\theta}}^{-1} \cdot \mathbf{x}_t | \mathbf{e}_i] \approx \frac{1}{N} \sum_k^N \mathbf{x}_{i,k}^T \cdot \mathbf{K}_{\boldsymbol{\theta}}^{-1} \cdot \mathbf{x}_{i,k}, \tag{7}$$

where $\{\mathbf{x}_{i,k}\}$ are samples generated by Gibbs sampling the probabilistic model given evidence $\mathbf{e}_i$. $N$ is empirically set to $10,000$ in our implementation.

**M-step**: We find $\boldsymbol{\theta}$ that maximizes the expected log-likelihood:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}^{(t)}}[L(\boldsymbol{\theta}; \{\mathbf{x}_i\}) | \{\mathbf{e}_i\}]. \tag{8}$$

There is no closed form solution for the optimization above, but the gradient of $E_{\boldsymbol{\theta}^{(t)}}[L(\boldsymbol{\theta}; \{\mathbf{x}_t\}) | \{\mathbf{e}_t\}]$ can be obtained analytically. Therefore, we employ a gradient-based optimization technique. In particular, let $\theta_j$ be a parameter in $\boldsymbol{\theta}$. The gradient of $E_{\boldsymbol{\theta}^{(t)}}[L(\boldsymbol{\theta}; \{\mathbf{x}_t\}) | \{\mathbf{e}_t\}]$ with respect to $\theta_i$ can be written as:

$$\frac{\partial E_{\boldsymbol{\theta}^{(t)}}[L(\boldsymbol{\theta};\{\mathbf{x}_i\})|\{\mathbf{e}_i\}]}{\partial\theta_j} = \frac{1}{2}\sum_i tr(\mathbf{K}_{\boldsymbol{\theta}} \cdot \frac{\partial \mathbf{K}_{\boldsymbol{\theta}}}{\partial\theta_j}) - \frac{1}{2N}\sum_i\sum_k \mathbf{x}_{i,k}^T \cdot \mathbf{K}_{\boldsymbol{\theta}}^{-1} \cdot \frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial\theta_j} \cdot \mathbf{K}_{\boldsymbol{\theta}}^{-1} \cdot \mathbf{x}_{i,k}. \tag{9}$$

To optimize for $\boldsymbol{\theta}$, we use a local search method for Gaussian process regression in the Gaussian Process and Machine Learning (GPML) Toolbox [RW13]. However, optimization is prone to being stuck with local optima since the objective function is non-convex. Consequently, we run the optimization from multiple initial states and choose parameters of the trial that yield maximum likelihood.

# 5 Additional Composition Examples

We show more results by our approach, the heuristic method, and the manual tool in Figure 3 and Figure 4.

# References

[CRHC06] B. Chun, D. Ryu, W. Hwang, and H. Cho. An automated procedure for word balloon placement in cinema comics. *LNCS*, 4292:576–585, 2006.

[RW13] C. Rasmussen and C. Williams. Gaussian processes for machine learning matlab code. `http://www.gaussianprocess.org/gpml/code/matlab/doc/`, 2013.

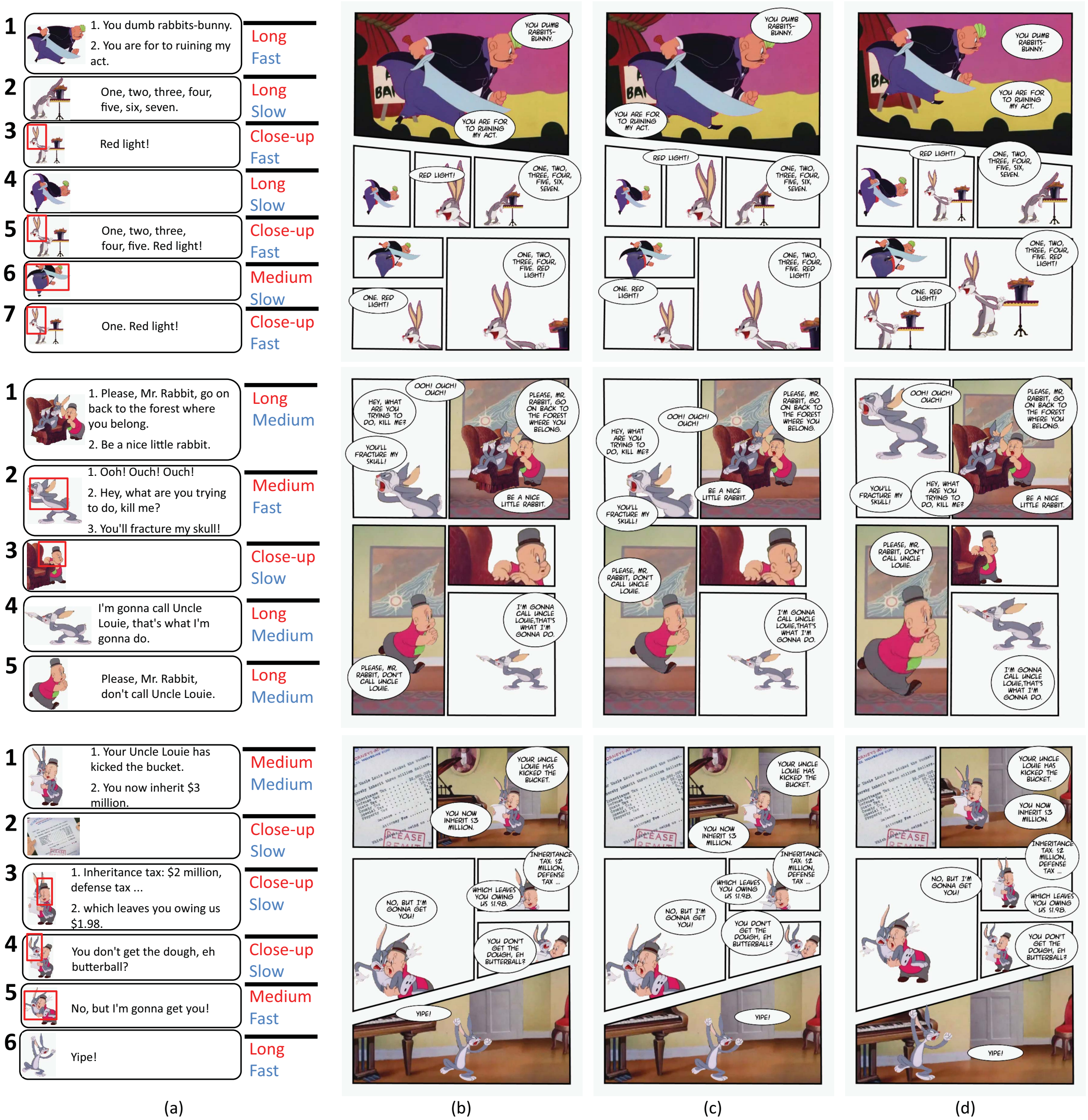
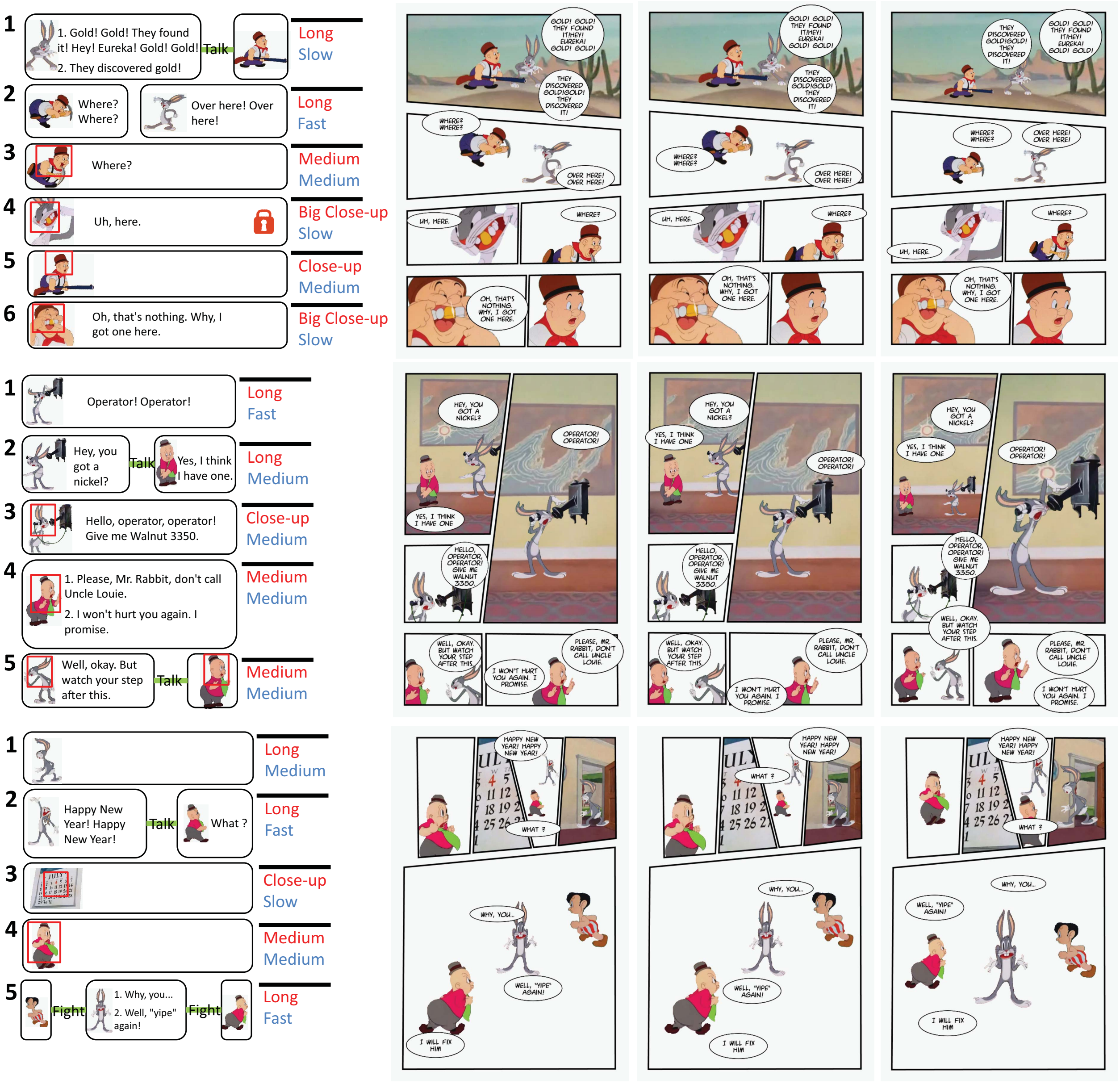

1
1. You dumb rabbits-bunny.
2. You are for to ruining my act.
Long
Fast
2
One, two, three, four, five, six, seven.
Long
Slow
3
Red light!
Close-up
Fast
4
Long
Slow
5
One, two, three, four, five. Red light!
Close-up
Fast
6
Medium
Slow
7
One. Red light!
Close-up
Fast
1
1. Please, Mr. Rabbit, go on back to the forest where you belong.
2. Be a nice little rabbit.
Long
Medium
2
1. Ooh! Ouch! Ouch!
2. Hey, what are you trying to do, kill me?
3. You'll fracture my skull!
Medium
Fast
3
Close-up
Slow
4
I'm gonna call Uncle Louie, that's what I'm gonna do.
Long
Medium
5
Please, Mr. Rabbit, don't call Uncle Louie.
Long
Medium
1
1. Your Uncle Louie has kicked the bucket.
2. You now inherit $3 million.
Medium
Medium
2
Close-up
Slow
3
1. Inheritance tax: $2 million, defense tax ...
2. which leaves you owing us $1.98.
Close-up
Slow
4
You don't get the dough, eh butterball?
Close-up
Slow
5
No, but I'm gonna get you!
Medium
Fast
6
Yipe!
Long
Fast
(a)
(b)
(c)
(d)

Figure 3: Compositions by our approach and the heuristic method. (a) Input storyboard. (b) Compositions by our approach. (d) Compositions by the heuristic method, with locations of subjects determined by our approach. (c) Compositions by participants using the manual tool. Input subjects and scripts at the first row are adapted from cartoon movies "*Bugs Bunny - Case of the Missing Hare*" (1942) in the public domain, while those at the second and third rows are from "*The Wabbit Who Came to Supper*" (1942) in the public domain.



(a) (b) (c) (d)

Figure 4: Compositions by our approach and the heuristic method. (a) Input storyboard. (b) Compositions by our approach. (d) Compositions by the heuristic method, with locations of subjects determined by our approach. (c) Compositions by participants using the manual tool. Input subjects and scripts at the first row are adapted from cartoon movies "*The Wacky Wabbit*" (1942) in the public domain, while those at the second and third rows are from "*The Wabbit Who Came to Supper*" (1942) in the public domain.