Supplementary Material for "VISIOCITY: A New Benchmarking Dataset and Evaluation Framework Towards Realistic Video Summarization"

1 Annotation Guidelines

As discussed in the main text, the ground truth in VISIOCITY is not *direct* in form of the user summaries, but *indirect* in form of concepts marked for each snippet (see main text for details). A group of 13 professional annotators were tasked to annotate videos (without listening to the audio) by marking all applicable keywords on a snippet/shot through a python GUI application developed by us for this task (Figure 1).

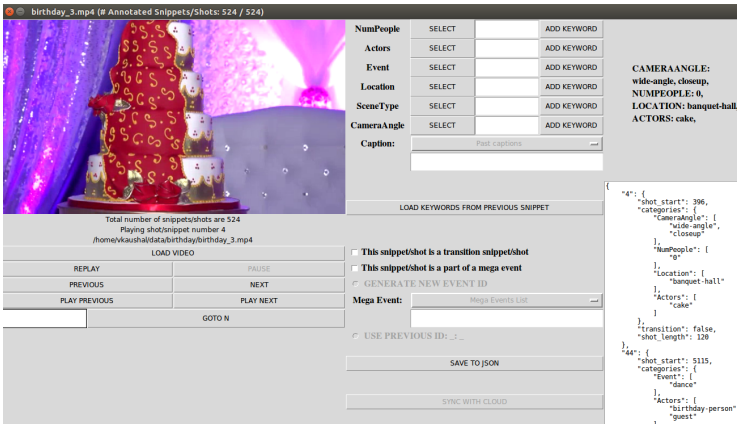

Figure 1: Annotation and visualization tool developed by us to create VISIOCITY

The guidelines and protocols were made as objective as possible, the annotators were trained through sample annotation tasks, and the annotation round was followed by two verification rounds where both precision (how accurate the annotations were) and recall (whether all events of interest and continuity information has been captured in the annotations) were verified by another set

of annotators. Whatever inconsistencies or inaccuracies were found and could be automatically detected, were included in our automatic sanity checks which were run on all annotations. **The detailed annotation guidelines that were provided to the annotators are available here.**

1.1 Instructions to setup the annotation tool

The tool can also be used for viewing the annotations or searching through the annotations given the annotation json file or as a summary viewer given the summary json file. The code is avilable from this git repository.

1. Pre-requisites: python3 and following python packages: tkinter, ffmpeg, opencv, pillow, imagetk, Pmw, bs4
2. **For annotation:** python tool.py soccer.json
3. **As annotation viewer:** python tool.py soccer.json vis
4. **Summary viewer:** GUI tool to view a summary, given its JSON. For example: python3 summaryViewer.py –video /data/soccer/soccer_1.mp4 –summary summary.json –annotation /data/soccer/soccer_1.json –configfile soccer.json

2 Accessibility and Long Term Availability of the dataset

VISIOCITY is and will continue to be hosted on the VISIOCITY project page. The videos, annotations and the human summaries are and will always be available for download from Google Drive accessible through the project page upon request.
Code for the annotation tool, the evaluation framework and some utilities is available at this git repository. The instructions to setup and execute are available on the project page.

3 Dataset organization and structure

1. **Videos** - there are six folders, each corresponding to a category. The videos are available as mp4 files under respective folders. Friends videos are available in avi format.
2. **Annotations** - the annotation for each video under each category is provided as a JSON file. For example, annotation for wedding_5.mp4 is in wedding_5.json. The annotation schema for each category is in the corresponding json file (say soccer.json) at this location.

3. **Summaries** - the human summaries used in the experiments are available as JSON files and are named as: userid_category_video-id_budget.json. The **JSON schema for a summary** is as follows:
   {
   "video_num_frames": number of frames in the video,
   "summary_num_frames": number of frames in the summary,
   "video_category": the category of the video, i.e. Friends or Surveillance, etc.,
   "mode": algorithm with which the summary has been produced,
   "video_name": file name of the video,
   "summary_num_snippets": number of snippets or shots in this summary,
   "num_snippets": number of snippets or shots in the video,
   "video_fps": fps of the video,
   "snippet_size": size of each snippet,
   "summary": binary vector of length video_num_frames indicating the summary frames as 1s
   "[snippet_id]": id of the snippet present in the summary
   }

4 Hourly wage paid to participants and the total amount spent on participant compensation:

1. Effort for training and sample annotation task on a 20 minutes video sample from each domain - 20 man hours
2. Effort for creating dense concept annotations for 67 videos - 280 man hours
3. Effort for creating human summaries for 12 videos - 24 man hours
4. Effort for two rounds of annotation verification - 210 man hours
5. Total man hours = 534 man hours
6. Wage paid to the participants = INR 700 per man hour
7. Total amount spent = INR 3,73,800, that is approximately, USD 5020

5 Declaration

To the best of our knowledge at the time of download, we have exercised caution to download only those videos that were available on YouTube with a Creative Commmons CC-BY (v3.0) License. As far as Friends videos are concerned, personal copy of purchased Friends videos were used.

As far as making the dataset available to others, we make them accessible through our project page upon request. Some videos may be subject to copyright. We don't own the copyright of those videos and only provide them for

non-commercial research purposes only. The annotation data provided by us can be used freely for research purposes.

6 Additional Results

6.1 Automatically generated ground truth summaries compared to human summaries and uniform and random summaries

In Table 4 of the main text we report aggregated results for all domains on AF1 measure. In Figure 3, we report aggregated results for our proposed performance measures as well for Soccer videos. Here we report min, mean and max for all performance measures (proposed, as well as maxF1 and avgF1) for all domains. The min, mean and max is across different budgets and different videos.

Technique	**AF1**	**MF1**	**IMP**	**MC**	**DT**	**DC**	**DSi**
human-min	18	27	39	28	41	45	68
human-mean	24	38	55	46	72	70	88
human-max	33	46	70	66	100	86	96
uniform-min	3	5	25	0	63	56	85
uniform-mean	5	9	31	7	89	66	90
uniform-max	8	14	35	10	100	77	93
random-min	5	9	28	10	20	33	82
random-mean	6	13	31	16	28	38	85
random-max	7	17	35	22	35	42	86
auto-min	21	32	85	55	60	82	77
auto-mean	25	41	87	69	76	85	81
auto-max	28	48	90	81	82	88	84

Table 1: Performance of Human and Auto summaries as compared to uniform and random summaries for Friends domain. The measures are reported in percentages.

6.2 Performance of different models on all domains in VISIOCITY

In Table 6 of the main text, we report the performance of different techniques including proposed VISIOCITY-SUM on Soccer and Friends videos. Here we report similar numbers for all other domains as well.

Technique	AF1	MF1	IMP	MC	DT	DC	DSi
human-min	22	32	40	36	29	57	56
human-mean	35	56	58	65	45	79	80
human-max	46	85	94	85	61	100	91
uniform-min	4	5	5	0	5	20	33
uniform-mean	6	8	12	9	12	49	55
uniform-max	9	13	18	17	19	81	71
random-min	3	6	9	0	8	34	40
random-mean	6	8	13	12	13	46	55
random-max	7	10	15	17	15	52	68
auto-min	24	33	73	64	73	93	81
auto-mean	31	40	80	85	82	99	88
auto-max	34	48	91	97	94	100	93

Table 2: Performance of Human and Auto summaries as compared to uniform and random summaries for **Surveillance** domain. The measures are reported in percentages.

Technique	AF1	MF1	IMP	MC	DT	DC	DSi
human-min	21	30	50	35	55	63	76
human-mean	30	45	56	55	75	84	85
human-max	39	53	66	69	100	95	89
uniform-min	2	5	21	14	21	30	78
uniform-mean	6	9	30	19	30	52	82
uniform-max	10	14	36	26	38	75	86
random-min	3	8	24	20	24	39	79
random-mean	5	9	30	22	30	51	81
random-max	7	10	36	26	37	65	84
auto-min	21	30	76	82	65	84	77
auto-mean	27	37	83	88	82	90	80
auto-max	31	44	88	91	88	95	83

Table 3: Performance of Human and Auto summaries as compared to uniform and random summaries for **Soccer** domain. The measures are reported in percentages.

Technique	AF1	MF1	IMP	MC	DT	DC	DSi
human-min	16	27	46	22	31	56	70
human-mean	21	31	56	38	48	70	83
human-max	29	41	76	55	82	83	92
uniform-min	3	6	40	0	46	44	78
uniform-mean	6	9	48	12	79	73	82
uniform-max	10	15	54	20	100	95	85
random-min	4	8	43	10	39	48	76
random-mean	6	10	48	16	47	57	78
random-max	7	12	53	20	55	65	81
auto-min	13	26	84	66	49	80	76
auto-mean	17	30	86	81	63	91	84
auto-max	19	31	88	88	78	96	90

Table 4: Performance of Human and Auto summaries as compared to uniform and random summaries for **Birthday** domain. The measures are reported in percentages.

Technique	AF1	MF1	IMP	MC	DT	DC	DSi
human-min	16	21	45	26	24	54	53
human-mean	21	39	57	39	46	69	76
human-max	30	59	74	63	100	79	89
uniform-min	4	6	33	0	69	61	73
uniform-mean	5	8	42	11	87	73	80
uniform-max	8	12	51	17	100	83	85
random-min	4	8	32	10	33	49	72
random-mean	5	9	42	18	40	54	78
random-max	6	10	49	24	48	58	83
auto-min	13	18	75	73	61	91	85
auto-mean	14	21	81	79	71	95	88
auto-max	16	22	86	87	79	97	89

Table 5: Performance of Human and Auto summaries as compared to uniform and random summaries for **Wedding** domain. The measures are reported in percentages.

Technique	**AF1**	**MF1**	**IMP**	**DT**	**DC**	**DSi**
human-min	8	14	26	37	74	35
human-mean	20	43	55	52	91	67
human-max	38	72	90	96	98	91
uniform-min	2	4	14	12	23	52
uniform-mean	7	9	49	29	45	60
uniform-max	11	13	84	52	80	69
random-min	2	5	14	14	26	38
random-mean	6	10	51	32	49	56
random-max	9	12	89	42	65	66
auto-min	11	24	69	48	81	82
auto-mean	25	43	86	78	93	96
auto-max	35	49	99	100	99	99

Table 6: Performance of Human and Auto summaries as compared to uniform and random summaries for **TechTalk** domain. The measures are reported in percentages.

Domain	Technique	AF1	MF1	IMP	MC	DT	DC	DSi
Soccer	Auto	59.3	93.3	83.2	84.3	82.6	85.9	76.2
	DR-DSN	2.8	8.9	23.7	20.3	23.2	30.4	**83.4**
	VASNET	28.4	43.4	**63**	49.3	**62.1**	**67.4**	75.2
	vsLSTM	**31.9**	**48.2**	**62.2**	**60.1**	**62**	**69.5**	76.5
	Ours	**32.6**	**50.3**	**64.2**	**62.6**	**63.4**	**72.2**	78.7
	Random	3.4	9.3	25.7	18.5	25.5	39.2	80.5
Friends	AUTO	66.3	96.9	87.8	84.6	80.3	89.8	83.1
	DR-DSN	4.3	9.4	19.1	6.9	**65.7**	51.5	**98.5**
	VASNET	**17**	**29.6**	**41**	**39.3**	49	**60.6**	86.7
	vsLSTM	15.5	27.2	**40.4**	**39.2**	**64.7**	59	91.1
	Ours	**17.4**	**31.2**	**42.5**	**40.5**	50.2	**64**	90.3
	Random	7.7	17.9	31.5	19.8	34.8	45.2	85.9
Surveillance	Auto	62.4	96.8	81.8	83.2	78.6	98	85.2
	DR-DSN	10	17.7	33.6	20.2	21.8	54.5	**57.2**
	VASNET	**19.4**	**31.4**	**39.5**	**42.6**	**28.4**	**65.4**	37.6
	vsLSTM	10.3	23.6	34.4	18.4	22.8	55.2	**58.4**
	Ours	**20.5**	**32.6**	**41.7**	**44.3**	**29.6**	**68.2**	38.5
	Random	3.9	8	16.6	12	15.3	49.4	69.4
TechTalk	Auto	64.7	91.5	79.8	-	80.5	88.4	94
	DR-DSN	13.5	22.5	49.3	-	24.8	29.9	35.2
	VASNET	**18.2**	**35.7**	52.1	-	**47.3**	**43.3**	**43.2**
	vsLSTM	15.1	32.2	**60.3**	-	38.8	35.3	41.7
	Ours	**18.7**	**37.5**	53.2	-	**50**	**45.8**	**45.5**
	Random	4.5	9.7	38.5	-	28	44	40.6
Birthday	Auto	67.3	97.2	89.7	88.6	68.1	90.8	81.3
	DR-DSN	8.1	14.2	54.7	14.1	**79.4**	63.6	**74.9**
	VASNET	21.6	37.6	50.1	30	36.2	47	48.7
	vsLSTM	**27.3**	**42.1**	**72.1**	**57.2**	59.6	**67.1**	73.6
	Ours	**28**	**44.3**	**74.8**	**60.3**	**62**	**69.5**	**77.6**
	Random	6.9	14.2	51.8	16.9	49.2	54.8	70.3
Wedding	Auto	55.4	94.4	83.9	74.7	67	88	85.7
	DR-DSN	4.2	8.9	40.7	14.4	**76.6**	**62**	**88.4**
	VASNET	4.5	14.4	46.5	22	44	52.7	84.9
	vsLSTM	**9**	**17.3**	**50.2**	**29.5**	50.1	56.9	80.7
	Ours	**9.4**	**17.9**	**52.8**	**30.3**	**51.8**	**58.6**	82.8
	Random	3.5	10	41.1	16.3	40.6	51.6	80

Table 7: Comparison of different techniques on VISIOCITY. TechTalk videos do not have MegaEvents