Learning Visual Prior via Generative Pre-Training

Anonymous Author(s)
Affiliation
Address
email

1 Appendix

1.1 Examples of Training Sequences

Here, we give some examples of various types of training sequences on different datasets:

Human Pose (COCO):
key point; multiple instances; large; 1; 18; person; [a 190 120 b 266 146 c 318 143 d 385 232 e 338 269 f 214 150 g 0 0 h 0 0 i 312 280 j 365 296 k 359 420 l 258 283 m 194 344 n 301 383 o 197 100 p 181 103 q 234 84 r 0 0]

Human Pose (CrowdPose):
key point; multiple instances; large; 2; 14; person, person; [a 312 201 b 306 200 c 311 232 d 269 214 e 298 257 f 231 206 g 296 275 h 307 275 i 251 244 j 271 235 k 274 292 l 283 295 m 304 153 n 310 191] [a 179 247 b 165 245 c 164 313 d 160 315 e 221 316 f 207 279 g 155 343 h 144 366 i 242 337 j 240 367 k 210 431 l 300 418 m 172 176 n 177 227] key point; multiple instances; large; 2; 14; person, person; [a 240 178 b 304 168 c 228 239 d 0 0 e 261 236 f 0 0 g 251 296 h 289 296 i 0 0 j 0 0 k 0 0 l 0 0 m 261 92 n 272 156] [a 314 160 b 363 158 c 274 232 d 356 264 e 224 260 f 271 263 g 298 315 h 341 324 i 0 0 j 332 442 k 0 0 l 0 0 m 287 64 n 333 133]

Instance Mask:
mask; multiple instances; medium; 1; 0; clock; [m0 224 291 m1 226 299 m2 227 306 m3 228 313 m4 233 320 m5 238 325 m6 245 329 m7 252 332 m8 259 334 m9 266 335 m10 274 333 m11 281 330 m12 288 327 m13 293 323 m14 299 318 m15 303 312 m16 305 305 m17 307 298 m18 310 291 m19 308 284 m20 307 276 m21 303 269 m22 299 263 m23 295 257 m24 288 254 m25 280 251 m26 273 250 m27 266 249 m28 259 249 m29 252 251 m30 246 256 m31 240 260 m32 235 265 m33 229 270 m34 227 277 m35 225 284]

Object Centric Bounding-Box:
box; object centric; large; 1; 0; castle; [xmin 236 ymin 142 xmax 413 ymax 232]

1.2 Implementation Details

All experimental evaluations were conducted on eight NVIDIA Tesla V100-32GB GPUs using PyTorch. In order to include special words, we created a new vocabulary containing a total of 30,769 words based on a standard vocabulary. To optimize computational efficiency and memory utilization, we utilized the DeepSpeed framework. To serialize visual locations, we first resized the long side of each image to a length of 512 pixels and then shifted the image content to the center by padding the short side to a length of 512 pixels. As a result, the number of bins m was set to 512. The flag of [Size] indicates the average area of all instances in the image and we set the flag according to the rule:

$$\begin{cases} \text{“small”} & \text{average area} < 32^2 \\ \text{“medium”} & 32^2 \leq \text{average area} < 96^2 \\ \text{“large”} & \text{average area} \geq 96^2 \end{cases}.$$

We omitted person instances with fewer than five keypoints. To enable continuous generation, we designed and trained models based on the prompt format (b). Specifically, VISORGPT† (a&b) and VISORGPT (a&b) were trained using the same number of sequences as VISORGPT† (a) and VISORGPT (a), respectively. The only difference is that we randomly utilized prompt format (a) or (b) to construct each training sequence.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

During the evaluation stage, we set the maximum sequence length of our model (VISORGPT) to 256 tokens to ensure efficient inference. In the ablation studies, we added special words only to the [Coordinate] term, and we reported the average KL divergence between the location and shape priors learned by VISORGPT and those in the real world. Since training large-scale language models is time- and resource-consuming, we trained only three types of VISORGPT with respect to GPT-2 (base, medium, large) with a maximum token length of 256 in 50,000 iterations on COCO (Box) data.

1.3 Evaluation Details

To estimate discrete visual prior from VISORGPT, we infer a series of sequences via prompting as below:

Code in Python:

```python
f"box; multiple instances; random.choice(['small', 'medium', 'large']);
random.randint(2, 10); 0; category name,"
```

To ensure that each category in a given dataset is sufficiently represented in the sequence data used for estimating the visual prior, we specify a minimum number of sequences in which each category must appear. Table 1 provides an overview of the predicted sequences that are used for evaluation.

Table 1: Details about the predicted sequences for evaluation.

Datasets	#Categories	#Predicted Seq.	Min #Seq. Per Category
Open Images (Box)	600	48,000	~80
Objects365 (Box)	365	29,200	~80
COCO (Box)	80	6,400	~80

In our study, we adopt the Kullback-Leibler divergence to quantify the similarity between two given discrete distributions. Specifically, let p and q denote the estimated probabilistic priors derived from the real-world data and the VISORGPT, respectively. The degree of similarity between these two distributions can be computed as:

$$\mathrm{KL}(p||q) = p\log(p/q). \tag{1}$$

1.4 Visualization

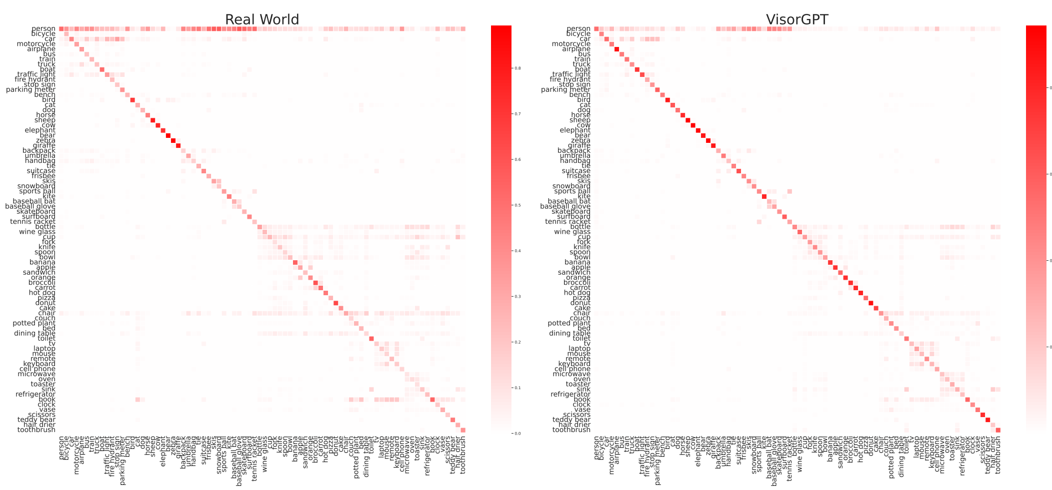


Figure 1: Relation among 80 categories on COCO.

Relation Prior of COCO. Fig. 1 illustrates the comparison between the real and learned relation prior among 80 categories on the COCO dataset. As can be observed, there is a high degree of similarity between the two relation matrices.

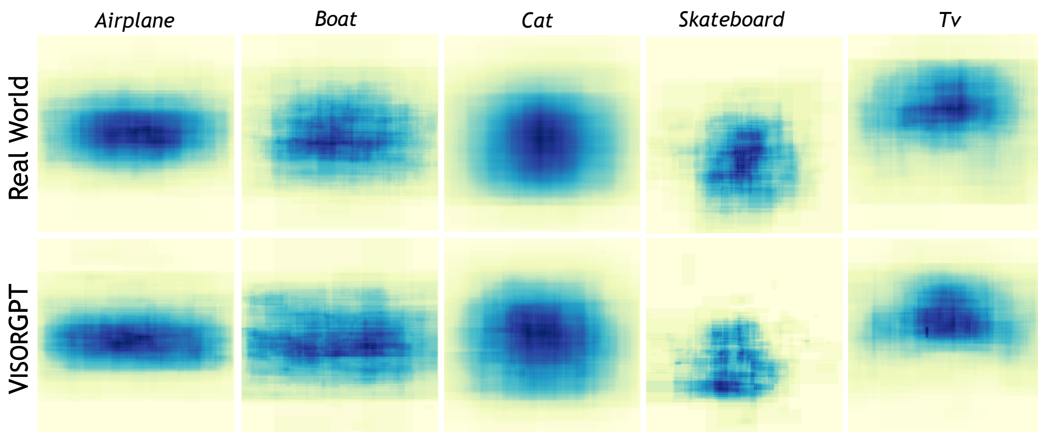
More Visual Comparison. We provide more comparison of visual prior between the real world and one learned by our VISORGPT and failure cases on COCO dataset in Fig. 2.

Continuous Generation. Fig. 3 presents a set of examples showcasing continuous generation based on the current scene. Notably, in each row, the proposed VISORGPT is able to successfully complete a scene that involves many individuals annotated with 14/18 keypoints or objects with bounding boxes, based on the information provided in the corresponding scene depicted in the previous columns.

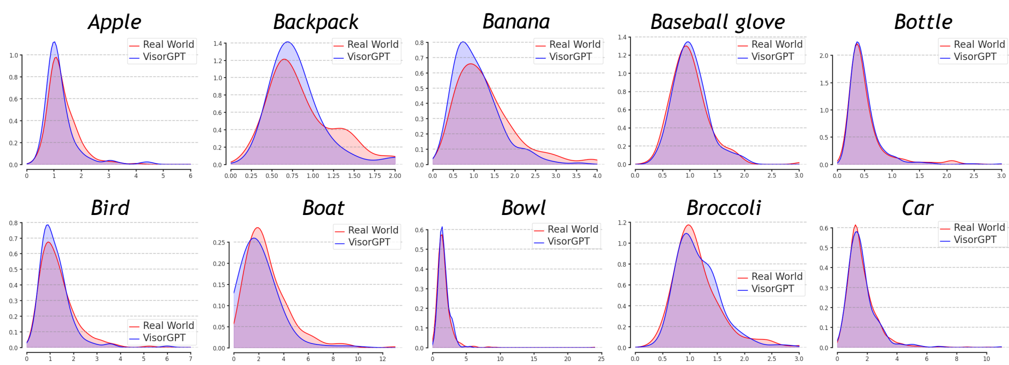
Figs. 4 and 5 present more visualization results.

1.5 Broader Impact

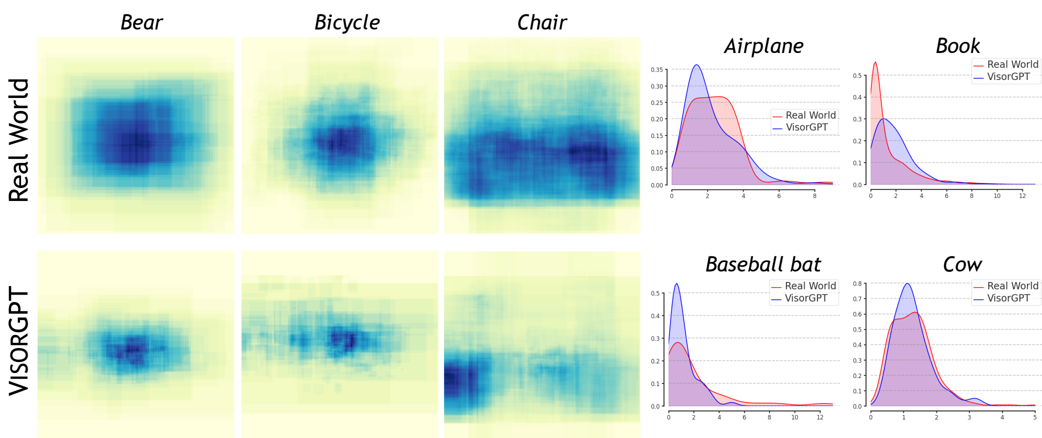
One of the key advantages of our VISORGPT is that it has learned the visual prior that can be used to sample various types of sequences. This allows for a high degree of customization in terms of the spatial conditions of bounding boxes, human poses, and instance masks, from many aspects such as object size, number of instances, and classes. In this way, the generated spatial conditions can be used to continuously synthesize paired image-box/pose/mask data such that we can potentially train more generalized visual intelligence models that are capable of handling a wider range of scenarios.



(a) Comparison of location prior

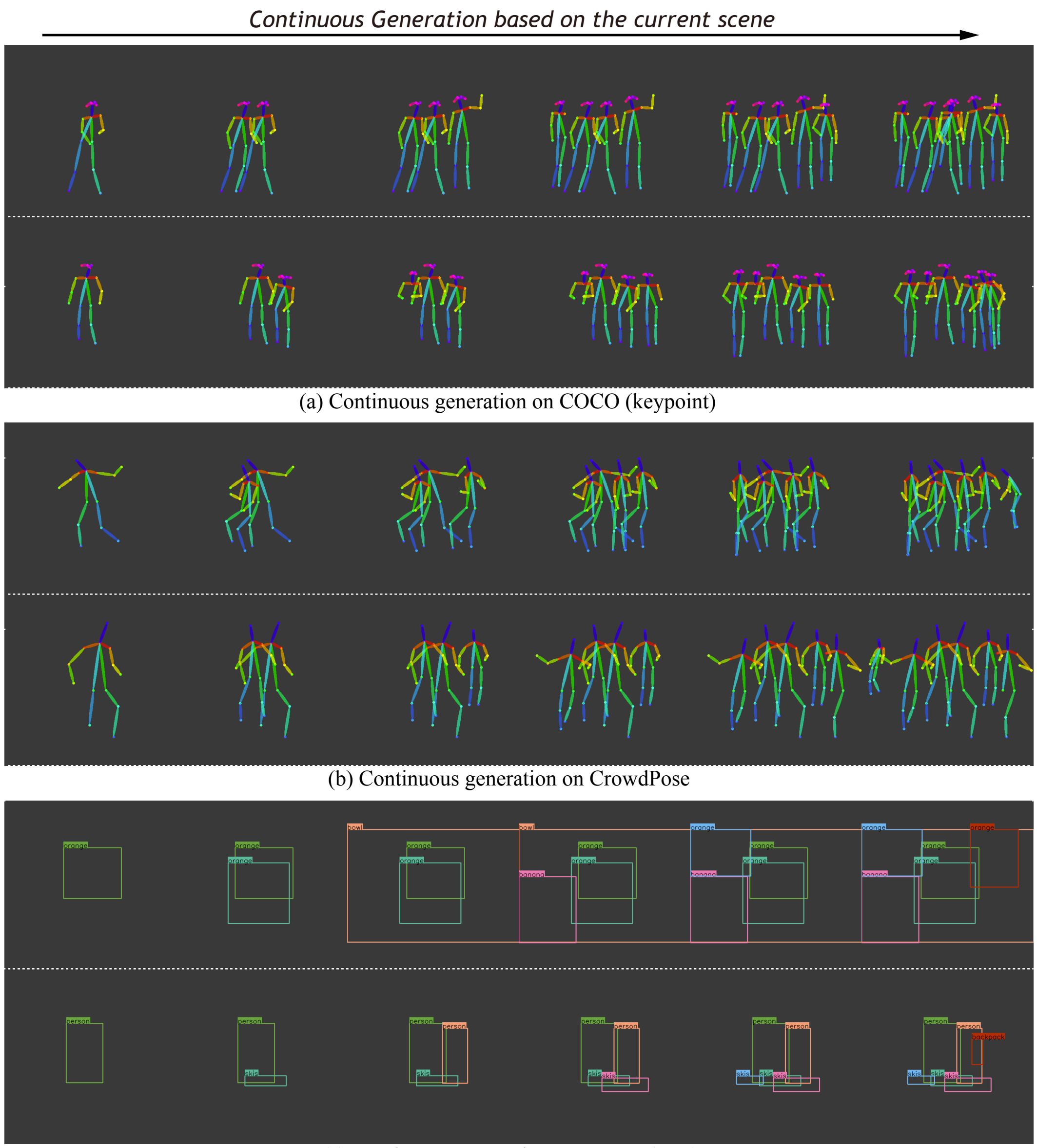


(b) Comparison of shape prior



(b) Failure cases

Figure 2: Comparison of visual prior between the real world and one learned by VISORGPT on COCO dataset.



(a) Continuous generation on COCO (keypoint)

(b) Continuous generation on CrowdPose

(c) Continuous generation on COCO (Box)

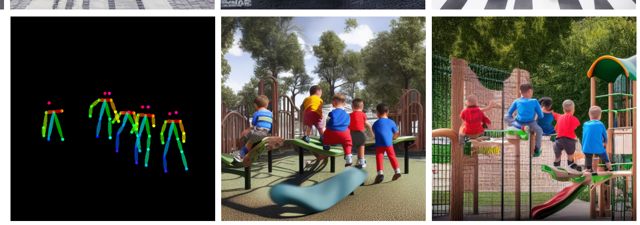
Figure 3: Examples of continual generation.

Input prompt:
key point; multiple instances; large; 13; 18; person,
Output sequence:
key point; multiple instances; large; 13; 18; person, person, person, person, person, person, person, person, person, person, person, person, person; [a 444 184 b 438 197 c 450 196 d 463 192 e 444 179 f 427 197 g 0 0 h 0 0 i 447 236 j 450 268 k 452 302 l 431 237 m 434 269 n 430 299 o 445 179 p 440 179 q 0 0 r 432 180] [a 232 168 b 223 178 c 238 174 d 267 150 e 256 142 f 208 182 g 183 187 h 149 192 i 230 246 j 262 291 k 281 345 l 204 248 m 208 298 n 187 338 o 236 163 p 228 164 q 0 0 r 217 163] [a 56 139 b 50 154 c 64 149 d 88 176 e 70 163 f 36 159 g 48 193 h 54 166 i 72 215 j 75 255 k 88 293 l 56 218 m 61 265 n 64 302 o 56 134 p 52 134 q 0 0 r 42 136] [a 148 145 b 137 162 c 150 162 d 0 0 e 0 0 f 124 161 g 122 188 h 129 216 i 150 225 j 0 0 k 0 0 l 132 228 m 145 264 n 0 0 o 150 142 p 142 143 q 0 0 r 132 141] [a 352 177 b 342 188 c 358 186 d 364 209 e 376 224 f 326 190 g 316 217 h 341 235 i 358 243 j 362 280 k 364 324 l 335 248 m 338 284 n 334 322 o 354 173 p 348 174 q 0 0 r 336 174] [a 0 0 b 248 31 c 13 136 d 28 196 e 0 0 f 0 0 g 0 0 h 0 0 i 1 233 j 5 316 k 7 383 l 0 0 m 0 0 n 0 0 o 0 0 p 0 0 q 0 0 r 0 0] [a 452 145 b 462 158 c 480 157 d 495 132 e 484 126 f 444 159 g 422 138 h 423 116 i 462 217 j 462 254 k 464 308 l 440 217 m 440 253 n 436 305 o 457 139 p 449 139 q 466 137 r 0 0] [a 204 162 b 201 185 c 219 186 d 0 0 e 0 0 f 183 184 g 178 212 h 197 228 i 217 237 j 0 0 k 0 0 l 192 240 m 0 0 n 0 0 o 208 159 p 201 159 q 212 162 r 194 162] [a 176 170 b 167 181 c 180 184 d 188 212 e 182 192 f 154 178 g 142 209 h 159 196 i 182 249 j 198 280 k 196 325 l 158 250 m 167 298 n 0 0 o 180 168 p 174 165 q 0 0 r 166 163] [a 382 188 b 379 202 c 396 202 d 410 202 e 394 171 f 363 202 g 0 0 h 0 0 i 388 249 j 391 295 k 386 342 l 360 250 m 364 294 n 368 340 o 387 186 p 379 183 q 390 189 r 369 186] [a 284 141 b 276 164 c 301 163 d 324 140 e 331 118 f 252 166 g 220 176 h 190 182 i 289 233 j 316 282 k 303 354 l 260 232 m 243 291 n 212 341 o 288 135 p 277 136 q 0 0 r 267 138] [a 0 0 b 402 170 c 397 172 d 0 0 e 0 0 f 407 169 g 416 197 h 0 0 i 397 215 j 0 0 k 0 0 l 405 213 m 400 257 n 392 289 o 0 0 p 0 0 q 0 0 r 412 156] [a 109 176 b 106 188 c 117 188 d 119 213 e 108 199 f 96 187 g 91 209 h 92 194 i 112 232 j 113 263 k 0 0 l 97 233 m 97 264 n 0 0 o 111 174 p 106 174 q 0 0 r 100 175]

Input prompt:
key point; multiple instances; large; 8; 18; person,
Output sequence:
key point ; multiple instances ; large ; 8 ; 18 ; person , person , person , person , person , person , person , person ; [a 88 184 b 78 222 c 40 227 d 9 314 e 67 315 f 115 217 g 0 0 h 0 0 i 65 390 j 0 0 k 0 0 l 118 379 m 0 0 n 0 0 o 73 172 p 100 171 q 55 181 r 112 172] [a 452 188 b 222 94 c 447 248 d 416 295 e 0 0 f 0 0 g 0 0 h 474 328 i 438 363 j 0 0 k 0 0 l 492 371 m 0 0 n 0 0 o 441 181 p 465 179 q 431 189 r 487 185] [a 0 0 b 0 0 c 0 0 d 0 0 e 0 0 f 0 0 g 0 0 h 0 0 i 0 0 j 0 0 k 0 0 l 421 268 m 423 302 n 426 325 o 0 0 p 0 0 q 0 0 r 0 0] [a 268 200 b 278 223 c 254 222 d 0 0 e 0 0 f 300 224 g 316 257 h 308 290 i 263 312 j 268 364 k 276 406 l 297 312 m 300 366 n 301 410 o 264 196 p 274 194 q 0 0 r 287 196] [a 168 181 b 166 221 c 133 225 d 121 270 e 160 273 f 201 218 g 228 258 h 236 273 i 147 313 j 165 376 k 171 434 l 188 308 m 188 371 n 196 440 o 160 174 p 175 174 q 148 182 r 184 182] [a 0 0 b 0 0 c 0 0 d 0 0 e 0 0 f 0 0 g 0 0 h 0 0 i 300 287 j 306 340 k 316 388 l 274 280 m 280 338 n 278 391 o 0 0 p 0 0 q 0 0 r 0 0] [a 383 216 b 401 238 c 390 238 d 0 0 e 0 0 f 413 237 g 428 274 h 408 298 i 388 303 j 0 0 k 0 0 l 409 303 m 411 362 n 419 407 o 379 212 p 387 210 q 0 0 r 403 210] [a 219 200 b 220 224 c 194 225 d 0 0 e 0 0 f 246 224 g 0 0 h 0 0 i 196 303 j 202 368 k 203 418 l 238 305 m 238 364 n 237 415 o 214 197 p 224 197 q 208 200 r 0 0]

Input prompt:
key point; multiple instances; large; 8; 18; person,
Output sequence:
key point ; multiple instances ; large ; 8 ; 18 ; person , person , person , person , person , person , person , person ; [a 88 184 b 78 222 c 40 227 d 9 314 e 67 315 f 115 217 g 0 0 h 0 0 i 65 390 j 0 0 k 0 0 l 118 379 m 0 0 n 0 0 o 73 172 p 100 171 q 55 181 r 112 172] [a 452 188 b 222 94 c 447 248 d 416 295 e 0 0 f 0 0 g 0 0 h 474 328 i 438 363 j 0 0 k 0 0 l 492 371 m 0 0 n 0 0 o 441 181 p 465 179 q 431 189 r 487 185] [a 0 0 b 0 0 c 0 0 d 0 0 e 0 0 f 0 0 g 0 0 h 0 0 i 0 0 j 0 0 k 0 0 l 421 268 m 423 302 n 426 325 o 0 0 p 0 0 q 0 0 r 0 0] [a 268 200 b 278 223 c 254 222 d 0 0 e 0 0 f 300 224 g 316 257 h 308 290 i 263 312 j 268 364 k 276 406 l 297 312 m 300 366 n 301 410 o 264 196 p 274 194 q 0 0 r 287 196] [a 168 181 b 166 221 c 133 225 d 121 270 e 160 273 f 201 218 g 228 258 h 236 273 i 147 313 j 165 376 k 171 434 l 188 308 m 188 371 n 196 440 o 160 174 p 175 174 q 148 182 r 184 182] [a 0 0 b 0 0 c 0 0 d 0 0 e 0 0 f 0 0 g 0 0 h 0 0 i 300 287 j 306 340 k 316 388 l 274 280 m 280 338 n 278 391 o 0 0 p 0 0 q 0 0 r 0 0] [a 383 216 b 401 238 c 390 238 d 0 0 e 0 0 f 413 237 g 428 274 h 408 298 i 388 303 j 0 0 k 0 0 l 409 303 m 411 362 n 419 407 o 379 212 p 387 210 q 0 0 r 403 210] [a 219 200 b 220 224 c 194 225 d 0 0 e 0 0 f 246 224 g 0 0 h 0 0 i 196 303 j 202 368 k 203 418 l 238 305 m 238 364 n 237 415 o 214 197 p 224 197 q 208 200 r 0 0]

Input prompt:
key point; multiple instances; large; 5; 18; person,
Output sequence:
key point ; multiple instances ; medium ; 5 ; 18 ; person , person , person , person , person ; [a 0 0 b 292 241 c 308 241 d 316 268 e 315 282 f 276 241 g 256 264 h 0 0 i 308 272 j 318 298 k 323 336 l 282 272 m 268 295 n 266 336 o 0 0 p 0 0 q 295 239 r 279 240] [a 0 0 b 236 209 c 251 209 d 262 238 e 270 262 f 221 209 g 202 225 h 199 249 i 248 256 j 249 303 k 0 0 l 224 256 m 217 300 n 0 0 o 0 0 p 0 0 q 247 192 r 236 192] [a 0 0 b 338 247 c 356 248 d 358 273 e 0 0 f 319 246 g 314 272 h 303 289 i 347 299 j 344 332 k 339 372 l 323 297 m 313 328 n 315 366 o 0 0 p 0 0 q 344 226 r 330 226] [a 0 0 b 108 238 c 128 236 d 0 0 e 0 0 f 89 240 g 83 298 h 82 278 i 130 307 j 0 0 k 0 0 l 95 307 m 0 0 n 0 0 o 0 0 p 0 0 q 0 0 r 96 216] [a 0 0 b 406 262 c 424 262 d 433 292 e 433 312 f 388 263 g 378 293 h 381 316 i 418 306 j 428 338 k 440 381 l 396 306 m 384 351 n 388 387 o 0 0 p 0 0 q 414 240 r 398 243]



Input prompt:
key point; multiple instances; large; 3; 18; person,
Output sequence:
key point ; multiple instances ; large ; 3 ; 18 ; person , person , person ; [a 164 220 b 167 266 c 203 279 d 0 0 e 212 324 f 132 252 g 95 302 h 112 338 i 184 352 j 0 0 k 0 0 l 127 344 m 0 0 n 0 0 o 176 210 p 158 210 q 192 215 r 146 214] [a 378 214 b 407 260 c 444 268 d 456 340 e 393 341 f 370 253 g 367 309 h 360 354 i 427 381 j 0 0 k 0 0 l 372 368 m 0 0 n 0 0 o 390 202 p 370 204 q 414 203 r 0 0] [a 296 238 b 317 281 c 355 289 d 355 349 e 299 364 f 278 273 g 263 324 h 274 353 i 336 391 j 0 0 k 0 0 l 282 382 m 0 0 n 0 0 o 307 232 p 292 230 q 329 237 r 0 0]

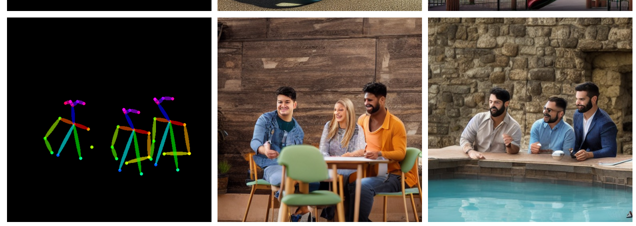


Figure 4: Examples of input prompts, output sequences, decoded results and synthetic images.

Input prompt:
box; multiple instances; large; 5; 0; dining table, knife, apple, apple, apple;
Output sequence:
box; multiple instances; large; 5; 0; dining table, knife, apple, apple, apple; [xmin 1 ymin 86 xmax 512 ymax 421] [xmin 0 ymin 85 xmax 100 ymax 388] [xmin 203 ymin 191 xmax 344 ymax 299] [xmin 366 ymin 254 xmax 472 ymax 355] [xmin 194 ymin 203 xmax 241 ymax 274]

Input prompt:
box; multiple instances; large; 5; 0; horse, person, person, horse, person;
Output sequence:
box; multiple instances; large; 5; 0; horse, person, person, horse, person; [xmin 1 ymin 164 xmax 219 ymax 414] [xmin 362 ymin 198 xmax 442 ymax 396] [xmin 234 ymin 169 xmax 316 ymax 327] [xmin 305 ymin 190 xmax 475 ymax 401] [xmin 350 ymin 203 xmax 383 ymax 297]

Input prompt:
box; multiple instances; large; 5; 0; horse, horse, horse, horse, horse;
Output sequence:
box; multiple instances; large; 5; 0; horse , horse , horse , horse , horse; [xmin 0 ymin 212 xmax 97 ymax 267] [xmin 0 ymin 166 xmax 59 ymax 237] [xmin 89 ymin 175 xmax 180 ymax 225] [xmin 235 ymin 144 xmax 512 ymax 421] [xmin 205 ymin 215 xmax 268 ymax 238]

Input prompt:
box; multiple instances; large; 7; 0; person, cup, donut, donut, donut, donut, dining table;
Output sequence:
box; multiple instances; large; 7; 0; person, cup, donut, donut, donut, donut, dining table; [xmin 0 ymin 64 xmax 509 ymax 158] [xmin 37 ymin 136 xmax 183 ymax 316] [xmin 235 ymin 144 xmax 352 ymax 249] [xmin 339 ymin 243 xmax 479 ymax 386] [xmin 212 ymin 239 xmax 357 ymax 378] [xmin 354 ymin 198 xmax 485 ymax 296] [xmin 1 ymin 64 xmax 511 ymax 442]

Input prompt:
box; multiple instances; large; 5; 0; boat, person, person, umbrella, tie;
Output sequence:
box; multiple instances; large; 5; 0; boat, person, person, umbrella, tie; [xmin 0 ymin 254 xmax 511 ymax 443] [xmin 245 ymin 146 xmax 381 ymax 391] [xmin 77 ymin 136 xmax 243 ymax 376] [xmin 340 ymin 98 xmax 510 ymax 193] [xmin 176 ymin 211 xmax 199 ymax 250]

Input prompt:
box; multiple instances; large; 5; 0; vase, chair, dining table, vase, chair;
Output sequence:
box; multiple instances; large; 5; 0; vase, chair, dining table, vase, chair; [xmin 189 ymin 249 xmax 305 ymax 411] [xmin 85 ymin 251 xmax 156 ymax 393] [xmin 85 ymin 347 xmax 423 ymax 506] [xmin 293 ymin 276 xmax 377 ymax 417] [xmin 345 ymin 243 xmax 427 ymax 346]

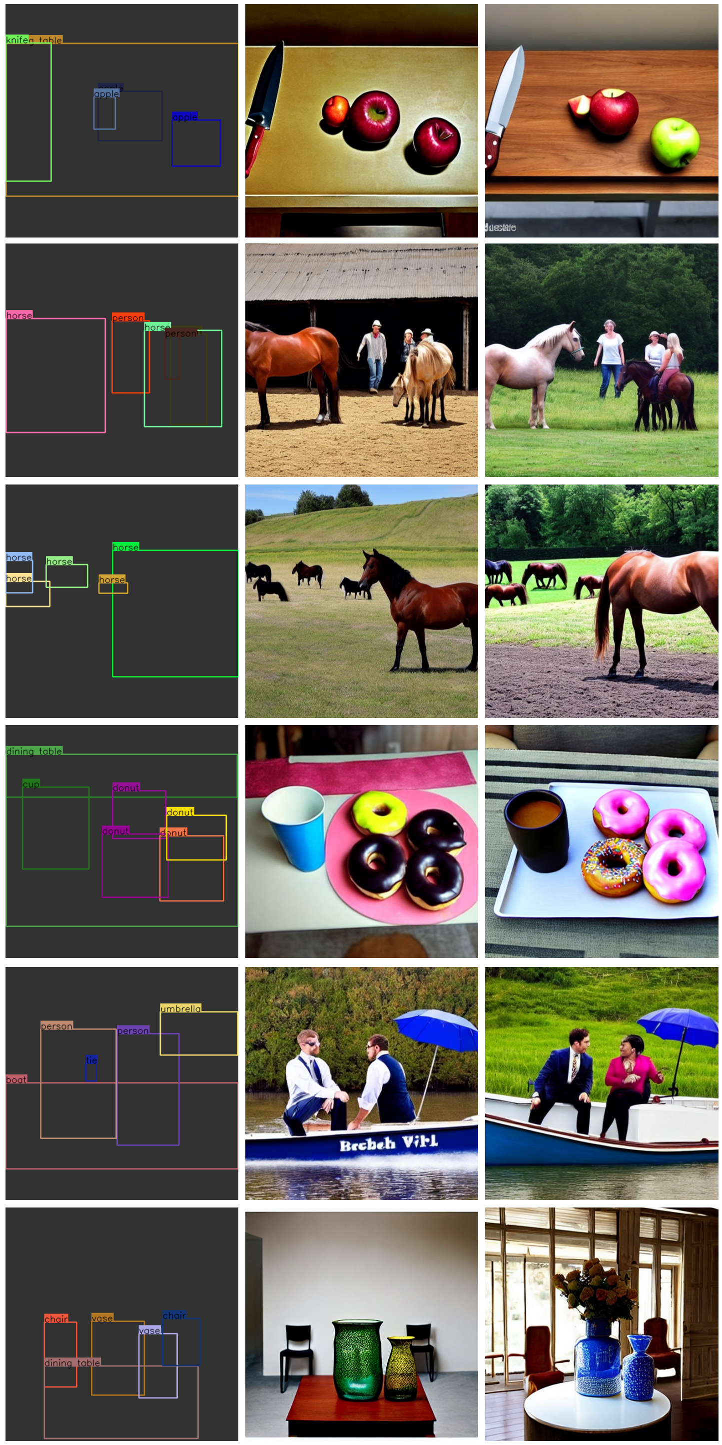


Figure 5: Examples of input prompts, output sequences, decoded results, and synthetic images.