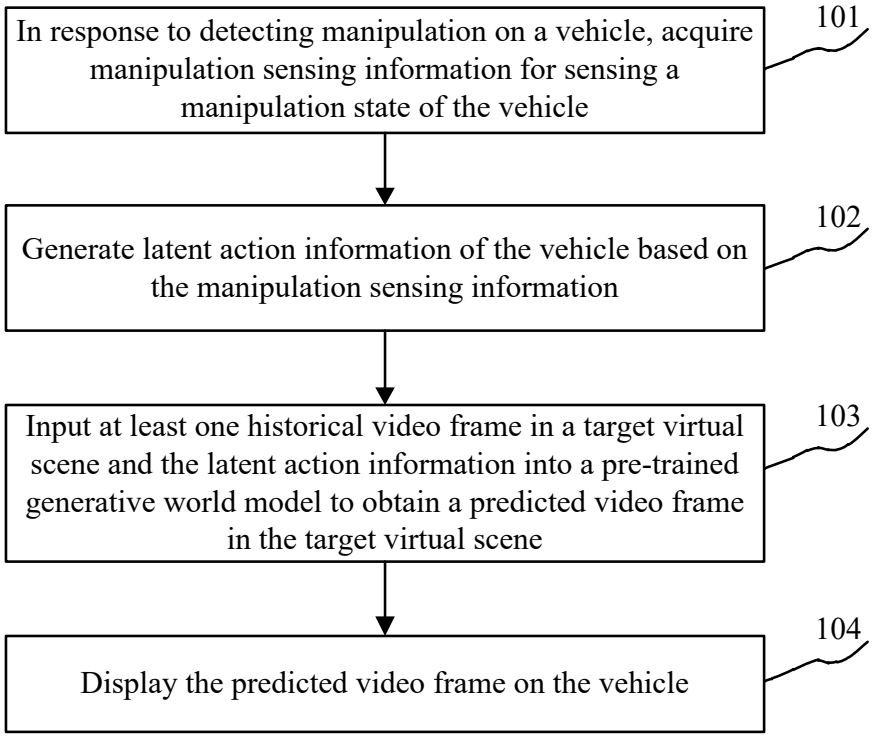
In response to detecting manipulation on a vehicle, acquire manipulation sensing information for sensing a manipulation state of the vehicle
101
Generate latent action information of the vehicle based on the manipulation sensing information
102
Input at least one historical video frame in a target virtual scene and the latent action information into a pre-trained generative world model to obtain a predicted video frame in the target virtual scene
103
Display the predicted video frame on the vehicle
104

FIG. 1

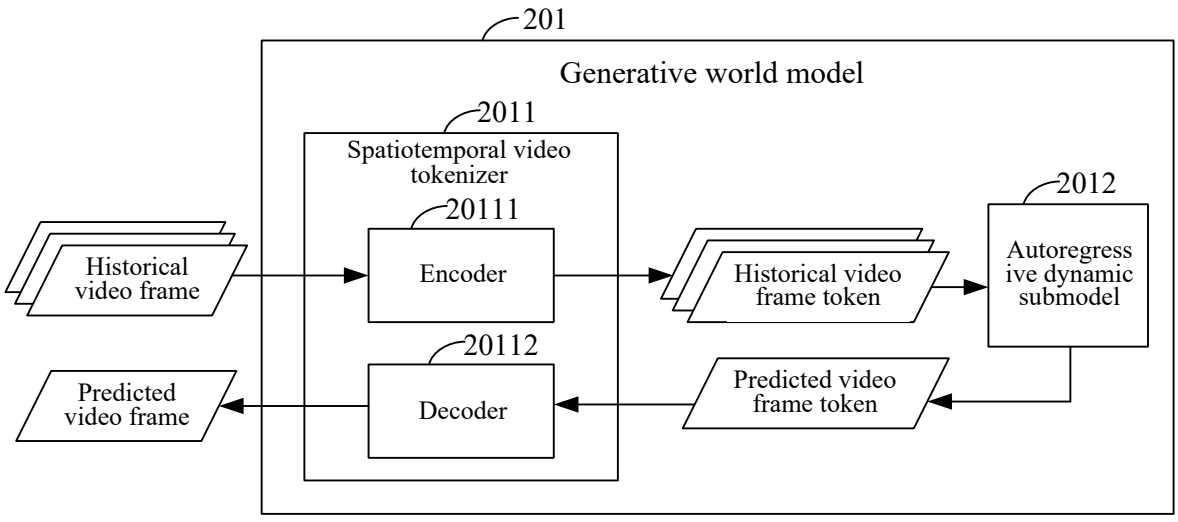
201
Generative world model
2011
Spatiotemporal video tokenizer
20111
Encoder
20112
Decoder
2012
Autoregressive dynamic submodel
Historical video frame
Predicted video frame
Historical video frame token
Predicted video frame token

FIG. 2

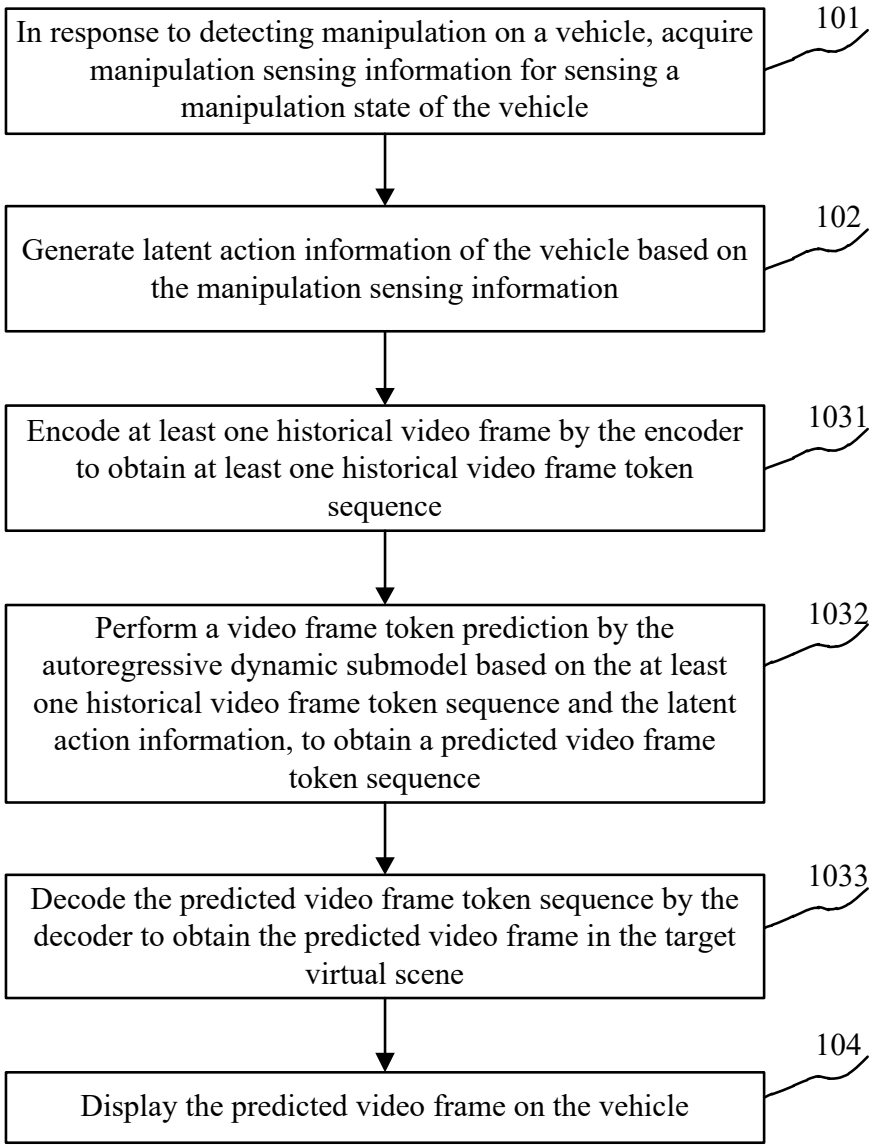
In response to detecting manipulation on a vehicle, acquire manipulation sensing information for sensing a manipulation state of the vehicle
101
Generate latent action information of the vehicle based on the manipulation sensing information
102
Encode at least one historical video frame by the encoder to obtain at least one historical video frame token sequence
1031
Perform a video frame token prediction by the autoregressive dynamic submodel based on the at least one historical video frame token sequence and the latent action information, to obtain a predicted video frame token sequence
1032
Decode the predicted video frame token sequence by the decoder to obtain the predicted video frame in the target virtual scene
1033
Display the predicted video frame on the vehicle
104

FIG. 3

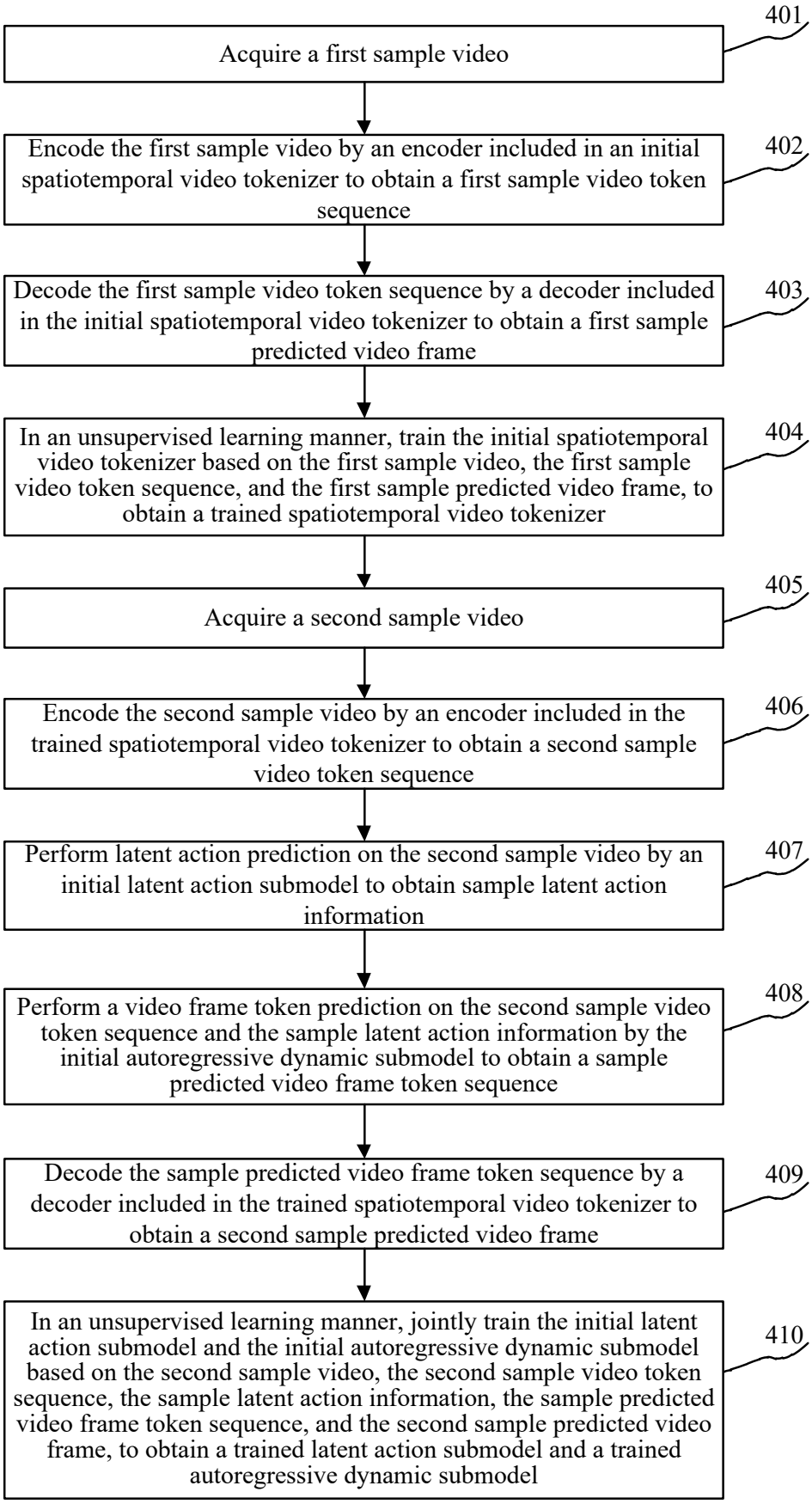
401
Acquire a first sample video
402
Encode the first sample video by an encoder included in an initial spatiotemporal video tokenizer to obtain a first sample video token sequence
403
Decode the first sample video token sequence by a decoder included in the initial spatiotemporal video tokenizer to obtain a first sample predicted video frame
404
In an unsupervised learning manner, train the initial spatiotemporal video tokenizer based on the first sample video, the first sample video token sequence, and the first sample predicted video frame, to obtain a trained spatiotemporal video tokenizer
405
Acquire a second sample video
406
Encode the second sample video by an encoder included in the trained spatiotemporal video tokenizer to obtain a second sample video token sequence
407
Perform latent action prediction on the second sample video by an initial latent action submodel to obtain sample latent action information
408
Perform a video frame token prediction on the second sample video token sequence and the sample latent action information by the initial autoregressive dynamic submodel to obtain a sample predicted video frame token sequence
409
Decode the sample predicted video frame token sequence by a decoder included in the trained spatiotemporal video tokenizer to obtain a second sample predicted video frame
410
In an unsupervised learning manner, jointly train the initial latent action submodel and the initial autoregressive dynamic submodel based on the second sample video, the second sample video token sequence, the sample latent action information, the sample predicted video frame token sequence, and the second sample predicted video frame, to obtain a trained latent action submodel and a trained autoregressive dynamic submodel

FIG. 4

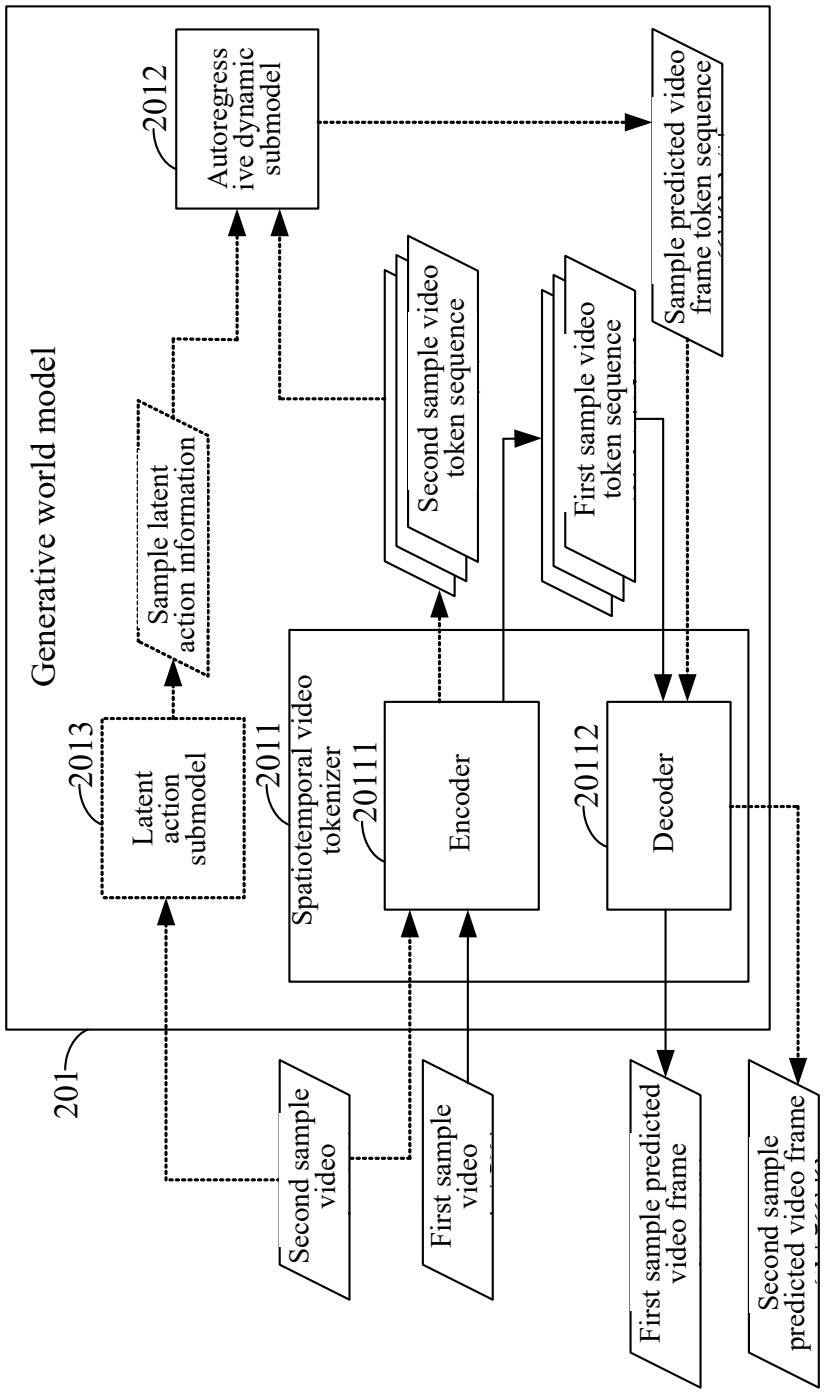
Generative world model
2012
Autoregressive dynamic submodel
Sample latent action information
Second sample video token sequence
First sample video token sequence
Sample predicted video frame token sequence
2013
Latent action submodel
2011
Spatiotemporal video tokenizer
20111
Encoder
20112
Decoder
201
Second sample video
First sample video
First sample predicted video frame
Second sample predicted video frame

FIG. 5

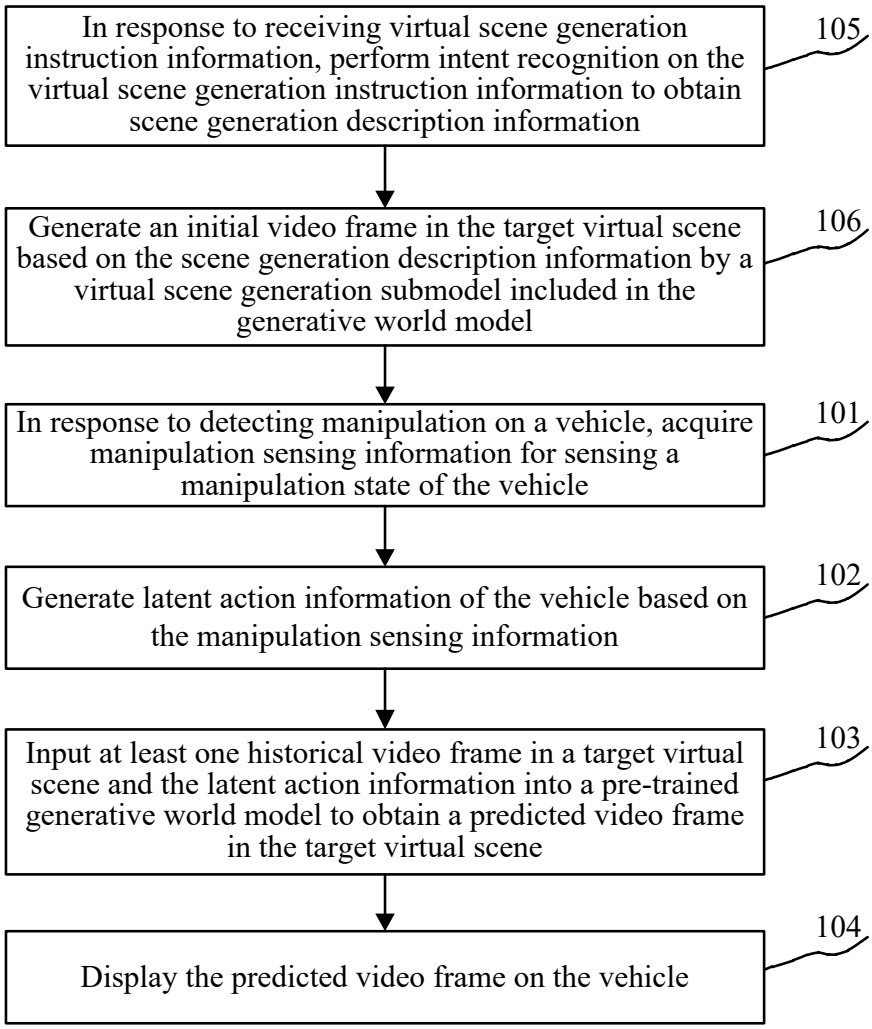
In response to receiving virtual scene generation instruction information, perform intent recognition on the virtual scene generation instruction information to obtain scene generation description information
105
Generate an initial video frame in the target virtual scene based on the scene generation description information by a virtual scene generation submodel included in the generative world model
106
In response to detecting manipulation on a vehicle, acquire manipulation sensing information for sensing a manipulation state of the vehicle
101
Generate latent action information of the vehicle based on the manipulation sensing information
102
Input at least one historical video frame in a target virtual scene and the latent action information into a pre-trained generative world model to obtain a predicted video frame in the target virtual scene
103
Display the predicted video frame on the vehicle
104

FIG. 6

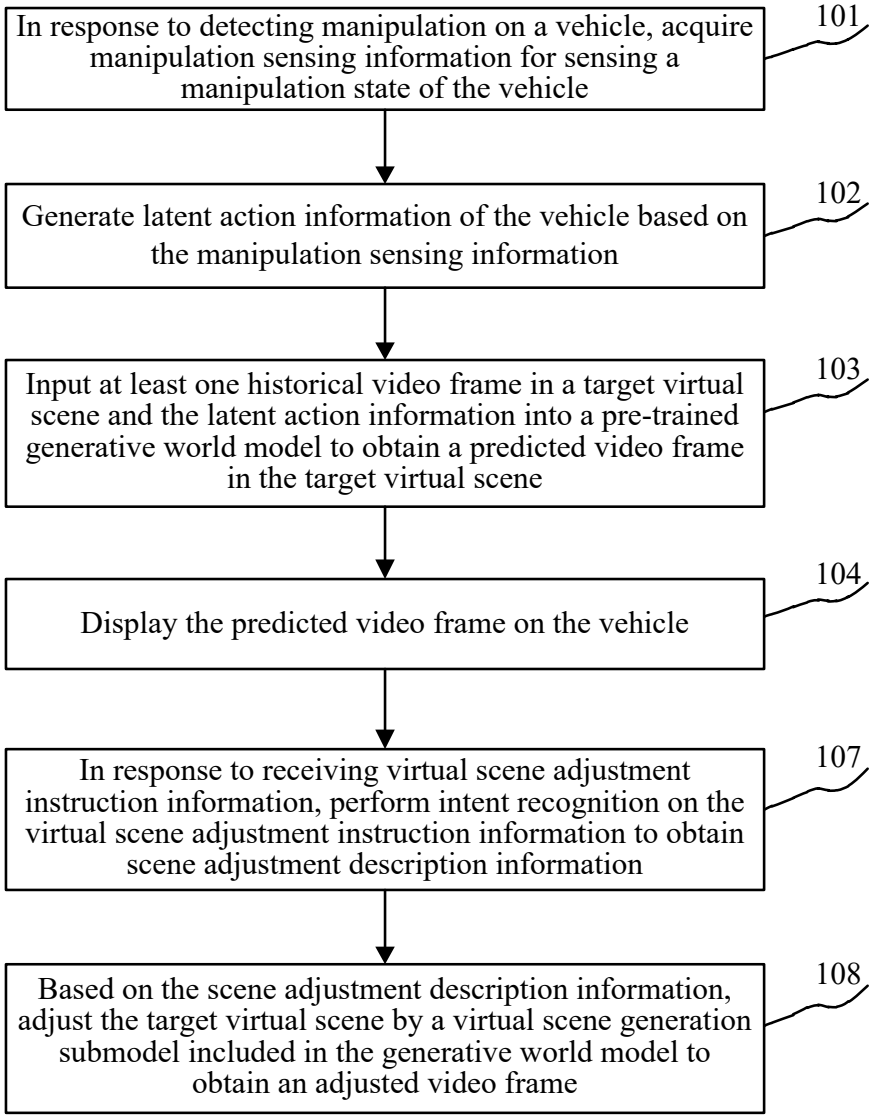
In response to detecting manipulation on a vehicle, acquire manipulation sensing information for sensing a manipulation state of the vehicle
101
Generate latent action information of the vehicle based on the manipulation sensing information
102
Input at least one historical video frame in a target virtual scene and the latent action information into a pre-trained generative world model to obtain a predicted video frame in the target virtual scene
103
Display the predicted video frame on the vehicle
104
In response to receiving virtual scene adjustment instruction information, perform intent recognition on the virtual scene adjustment instruction information to obtain scene adjustment description information
107
Based on the scene adjustment description information, adjust the target virtual scene by a virtual scene generation submodel included in the generative world model to obtain an adjusted video frame
108

FIG. 7

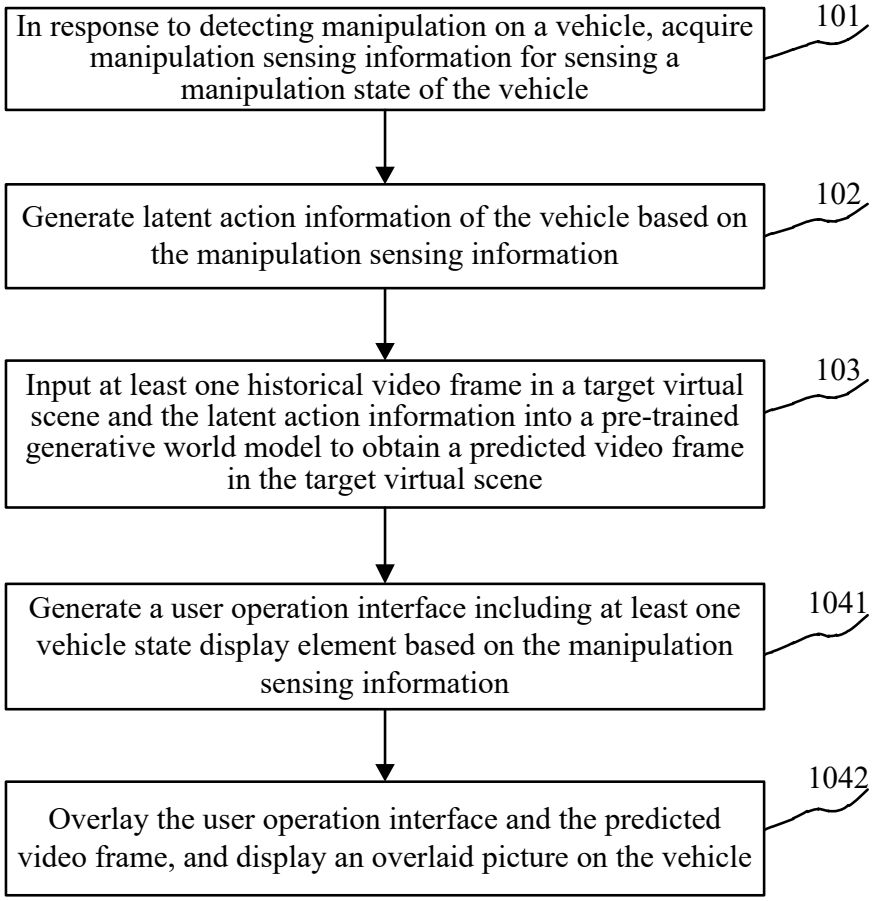
In response to detecting manipulation on a vehicle, acquire manipulation sensing information for sensing a manipulation state of the vehicle
101
Generate latent action information of the vehicle based on the manipulation sensing information
102
Input at least one historical video frame in a target virtual scene and the latent action information into a pre-trained generative world model to obtain a predicted video frame in the target virtual scene
103
Generate a user operation interface including at least one vehicle state display element based on the manipulation sensing information
1041
Overlay the user operation interface and the predicted video frame, and display an overlaid picture on the vehicle
1042

FIG. 8

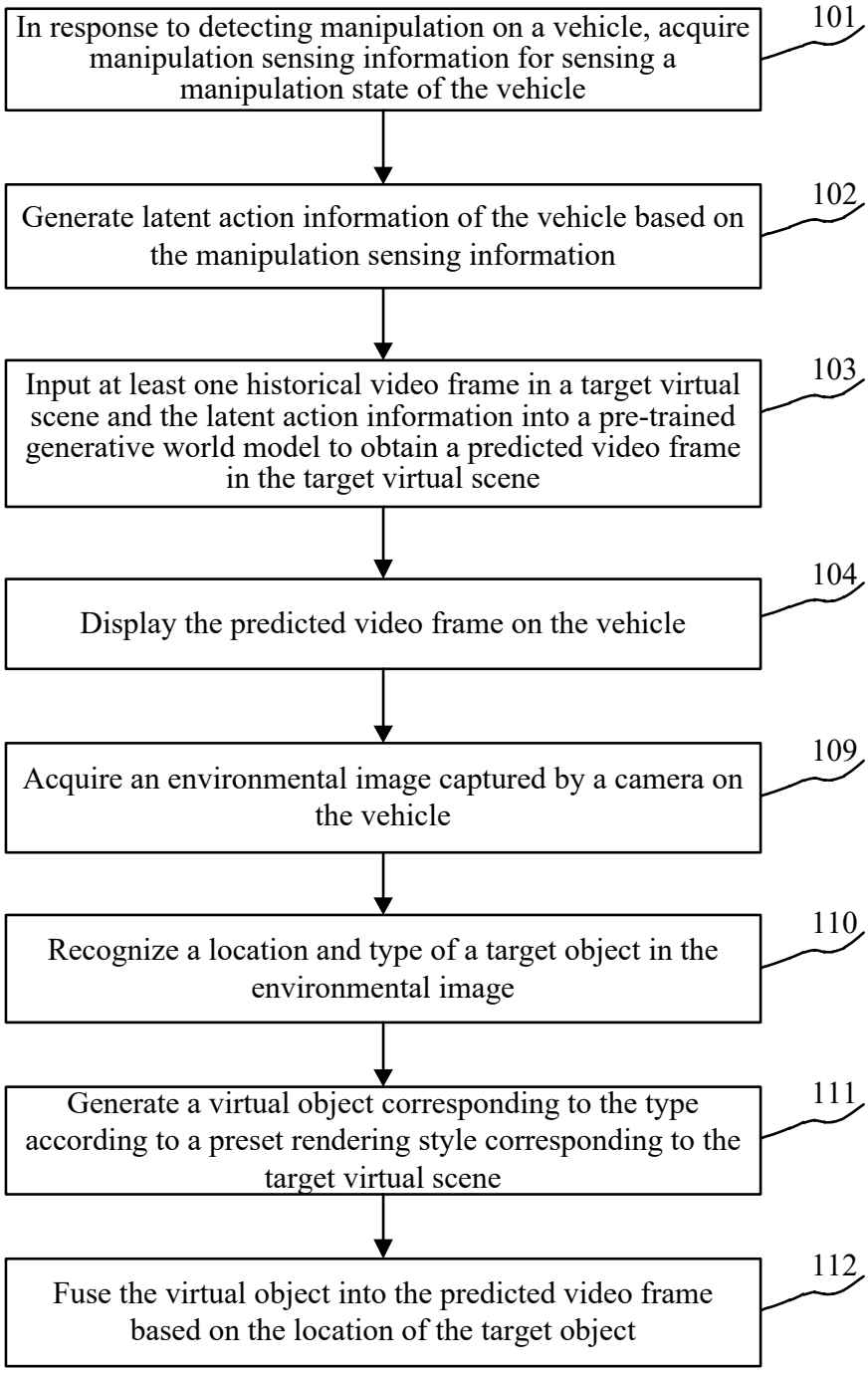
In response to detecting manipulation on a vehicle, acquire manipulation sensing information for sensing a manipulation state of the vehicle
101
Generate latent action information of the vehicle based on the manipulation sensing information
102
Input at least one historical video frame in a target virtual scene and the latent action information into a pre-trained generative world model to obtain a predicted video frame in the target virtual scene
103
Display the predicted video frame on the vehicle
104
Acquire an environmental image captured by a camera on the vehicle
109
Recognize a location and type of a target object in the environmental image
110
Generate a virtual object corresponding to the type according to a preset rendering style corresponding to the target virtual scene
111
Fuse the virtual object into the predicted video frame based on the location of the target object
112

FIG. 9

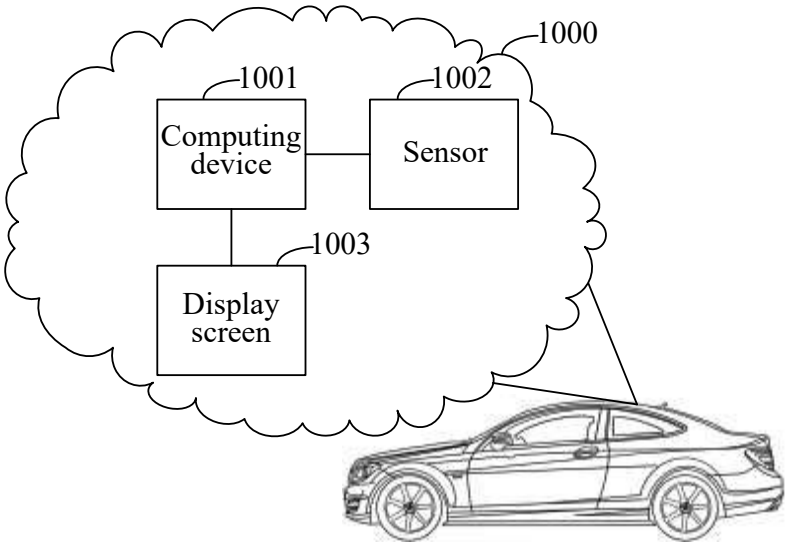
1000
1001
1002
Computing device
Sensor
1003
Display screen

FIG. 10

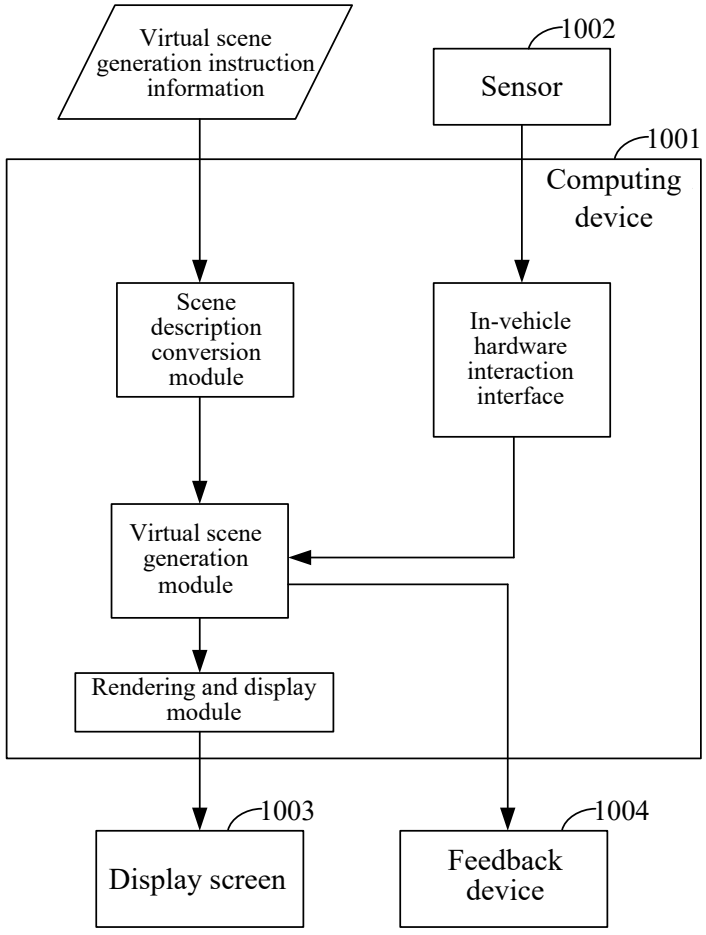
Virtual scene generation instruction information
1002
Sensor
1001
Computing device
Scene description conversion module
In-vehicle hardware interaction interface
Virtual scene generation module
Rendering and display module
1003
Display screen
1004
Feedback device

FIG. 11

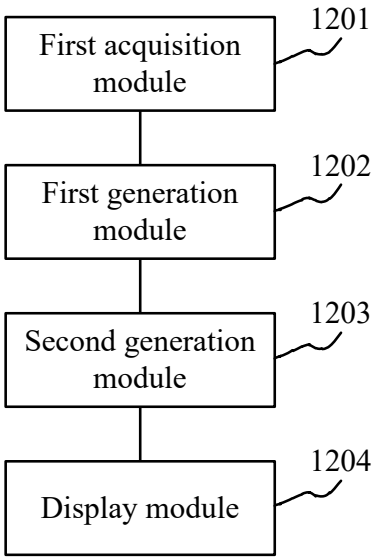
First acquisition module
1201
First generation module
1202
Second generation module
1203
Display module
1204

FIG. 12

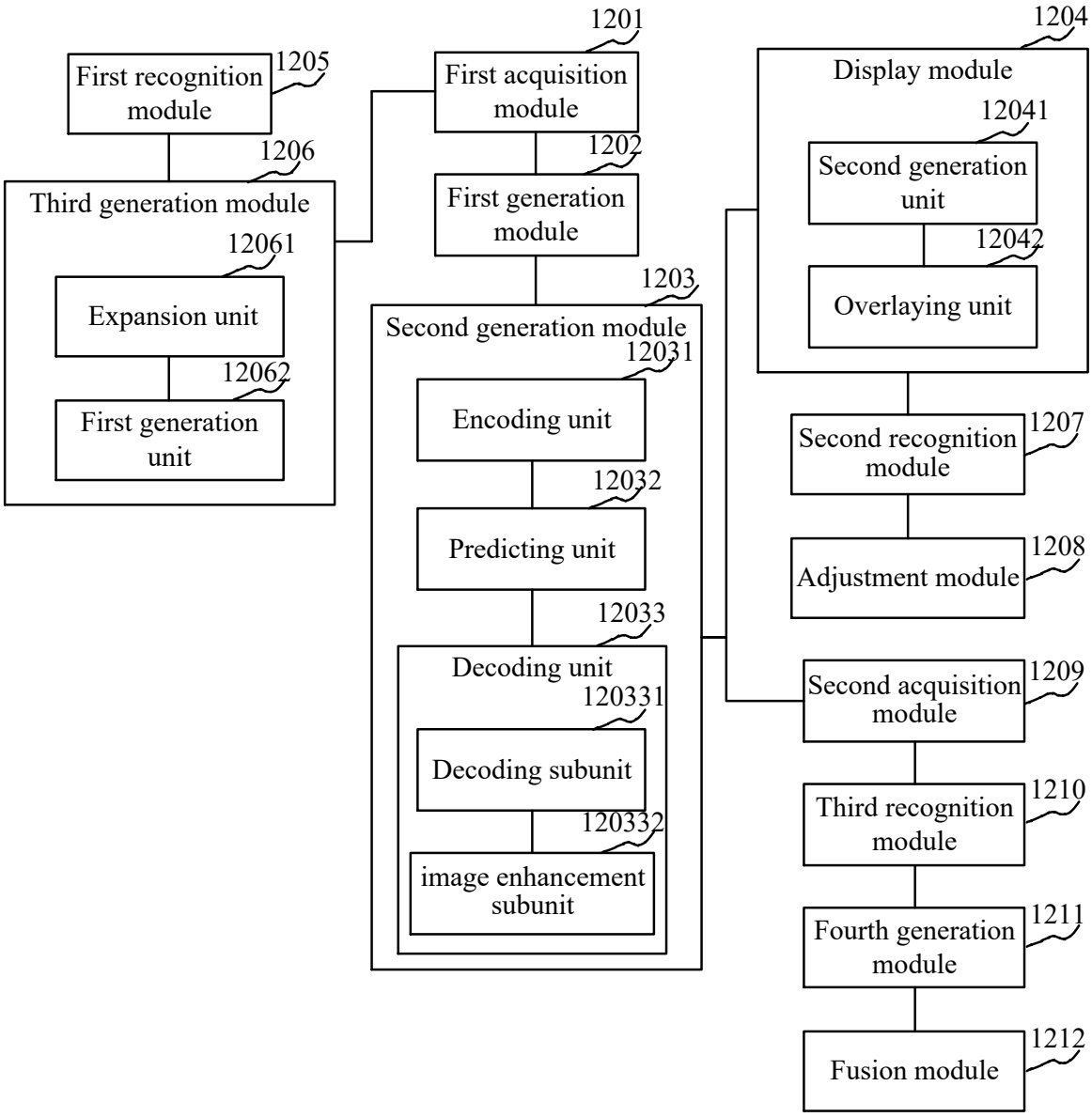
1201
First acquisition module
1202
First generation module
1203
Second generation module
12031
Encoding unit
12032
Predicting unit
12033
Decoding unit
120331
Decoding subunit
120332
image enhancement subunit
1204
Display module
12041
Second generation unit
12042
Overlaying unit
1205
First recognition module
1206
Third generation module
12061
Expansion unit
12062
First generation unit
1207
Second recognition module
1208
Adjustment module
1209
Second acquisition module
1210
Third recognition module
1211
Fourth generation module
1212
Fusion module

FIG. 13

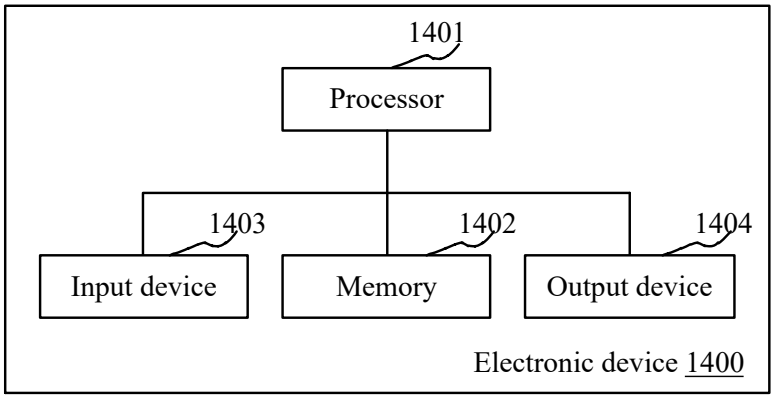
1401
Processor
1403
1402
1404
Input device
Memory
Output device
Electronic device 1400

FIG. 14